



Bayesian Methods in Pharmaceutical Development and Clinical Trial Design

Scott Berry

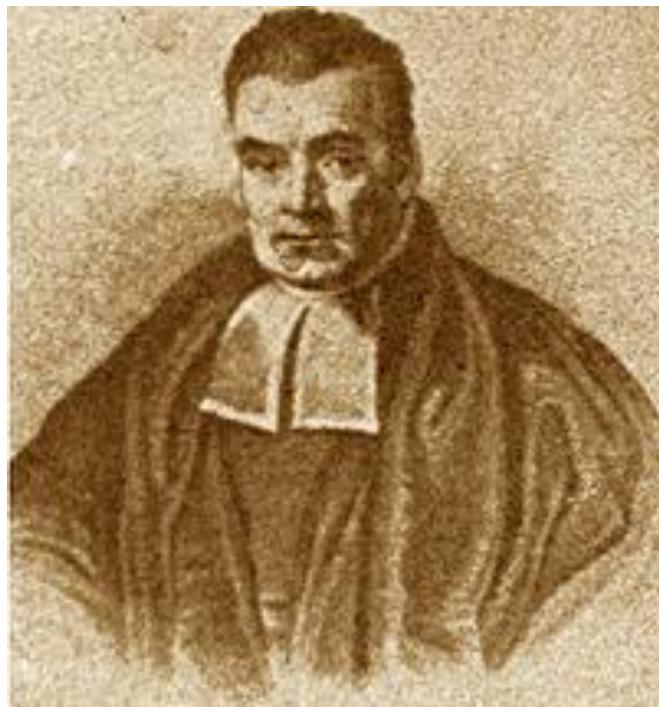
scott@berryconsultants.com

@statberry 

Outline

- Bayesian: What & Why
 - Bayesian Modeling
 - $\pi(\theta|x)$ and $f(x|\theta)$ really does matter
 - Prediction/Forecasting
 - Utilities
- The clash of philosophies in design
- Example using all of above!
- Discussion

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*

Bayes Theorem

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

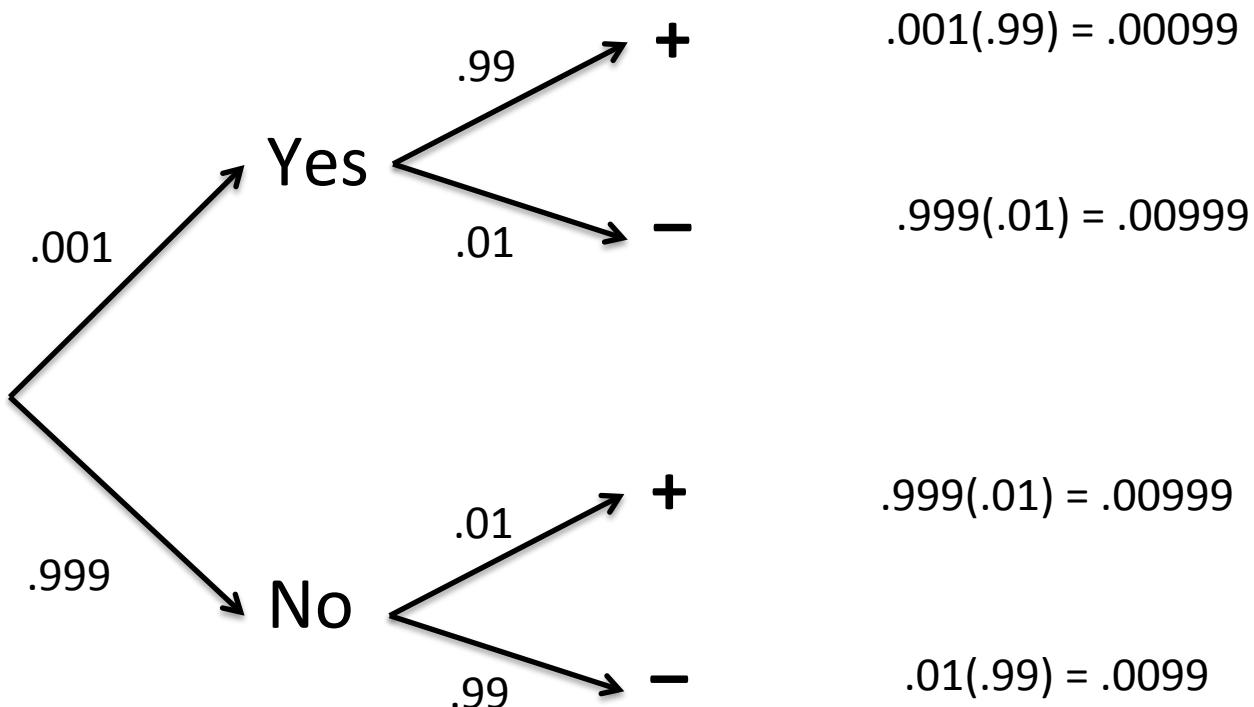
$$f(\theta | X) = \frac{f(x | \theta) \pi(\theta)}{\int f(x | \theta) \pi(\theta) d\theta}$$

Rare Disease Example

- Suppose a diagnostic test is 99% effective – meaning, $\Pr(+|\text{disease}) = 0.99$ and $\Pr(-|\text{no disease}) = 0.99$
- You test a random patient and get a positive (+) test, what is the probability the patient has the disease?
 - $\Pr(\text{Disease} | + \text{Test}) = ?$
 - We know $\Pr(+ \text{Test} | \text{Disease})$
- Suppose the incidence of the disease is 1 in 1000

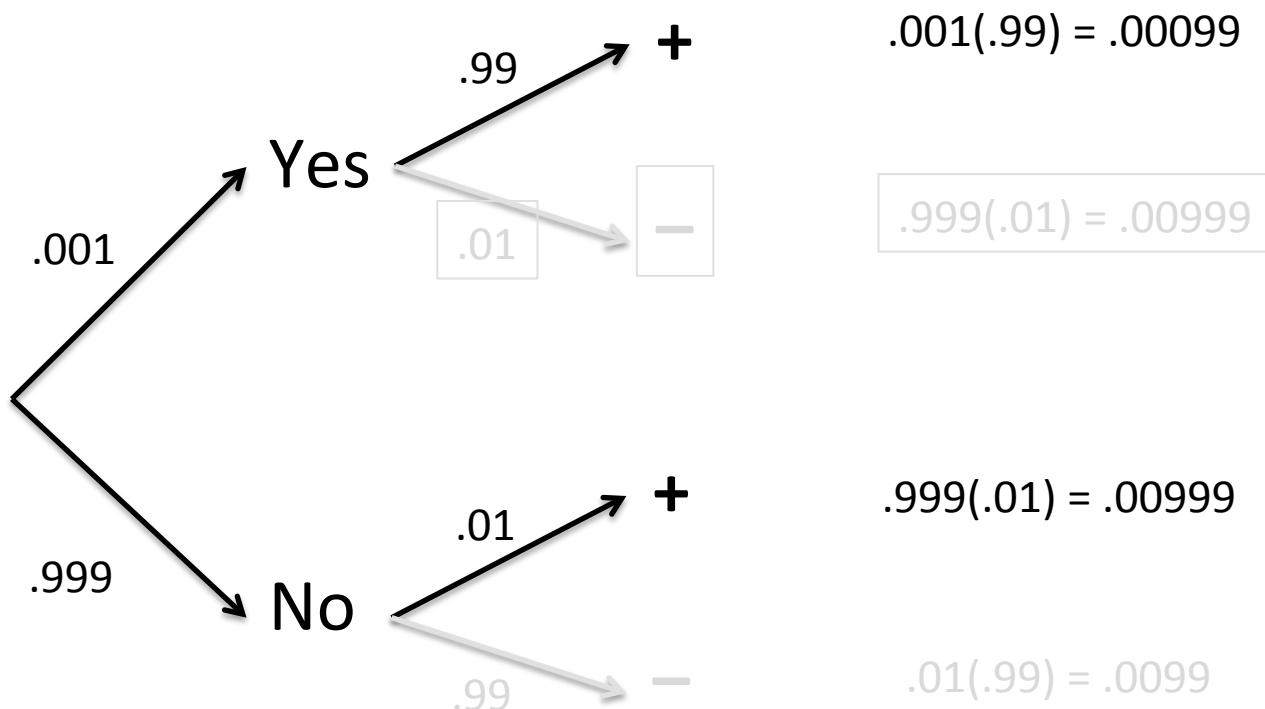
Rare Disease Example

- Possible outcomes:

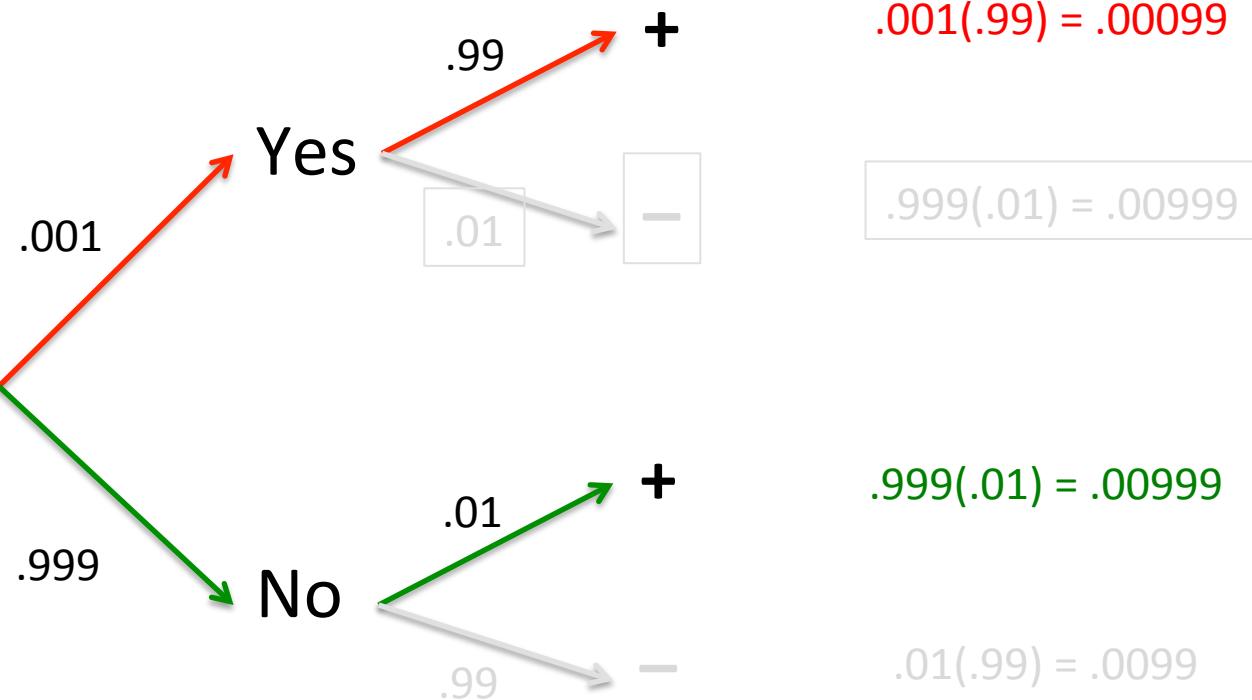


Rare Disease Example

- Those outcomes that could have happened



Bayes Theorem



$$\Pr(\text{Yes} \mid +) = \frac{\text{Top Right Box}}{\text{Top Right Box} + \text{Bottom Right Box}} = \frac{.00099}{.00099 + .00999} = .090$$

Compare P-Values/Posteriors

- Inspired by **Steve Ruberg** Example
- You have a bag of coins, mixed fair coins and 2-headed coins
 - Assume a null (H_0) of “fair coin”
 - Alternative (H_1) of “2-headed coin”
- Flip the coin independently n times...

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

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

Data/P-Values/Posteriors

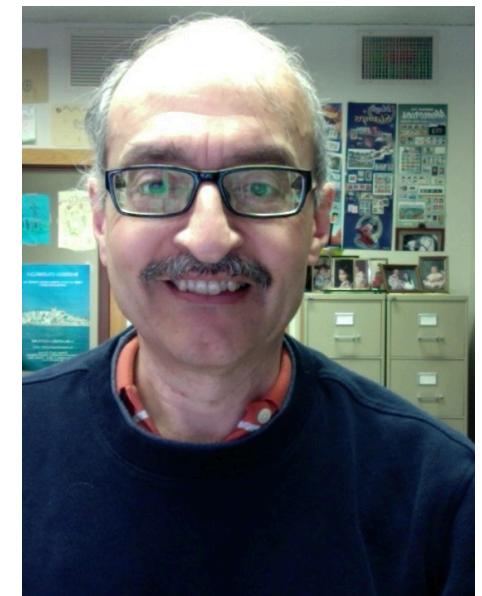
DATA	P-Value	Pr(Fair Coin)	
		50% each	0.001 2-headed
1/1	0.50	0.333	0.998
2/2	0.25	0.200	0.996
3/3	0.125	0.111	0.992
4/4	0.0625	0.0588	0.984
5/5	0.0312	0.0303	0.969
6/6	0.0156	0.0154	0.940
7/7	0.00781	0.00775	0.886
8/8	0.00391	0.00389	0.796
9/9	0.00195	0.00194	0.661
10/10	0.000977	0.000976	0.493
11/11	0.000488	0.000488	0.327
12/12	0.000244	0.000244	0.196
16/16	0.000015	0.000015	0.015

If you want to analyze data before you get it
be a frequentist, if you want to analyze
data you have be a Bayesian.

Paraphrased...

--Mark Schervish

(unsure if he is repeating others...)



Frequentist hypothesis testing

Bernoulli Observations, iid, $\text{pr}(S) = \pi$. Null of $\pi=0.5$

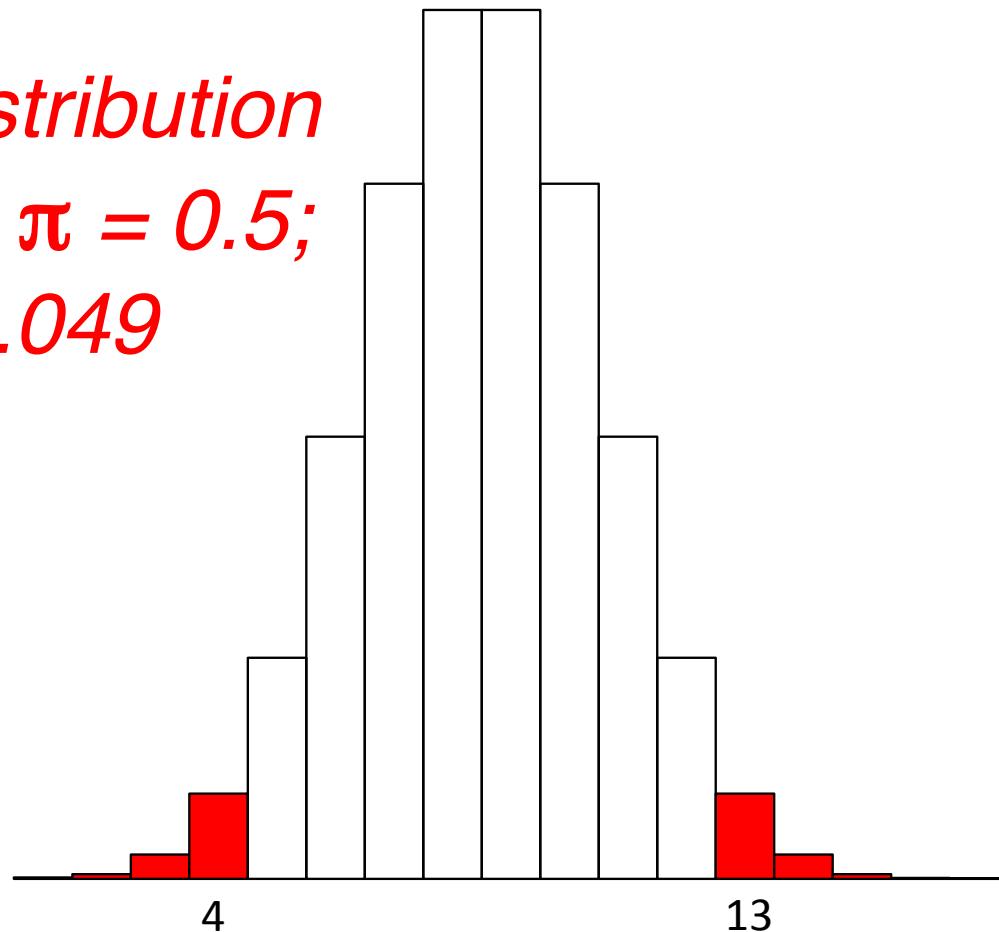
Four designs:

- (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"

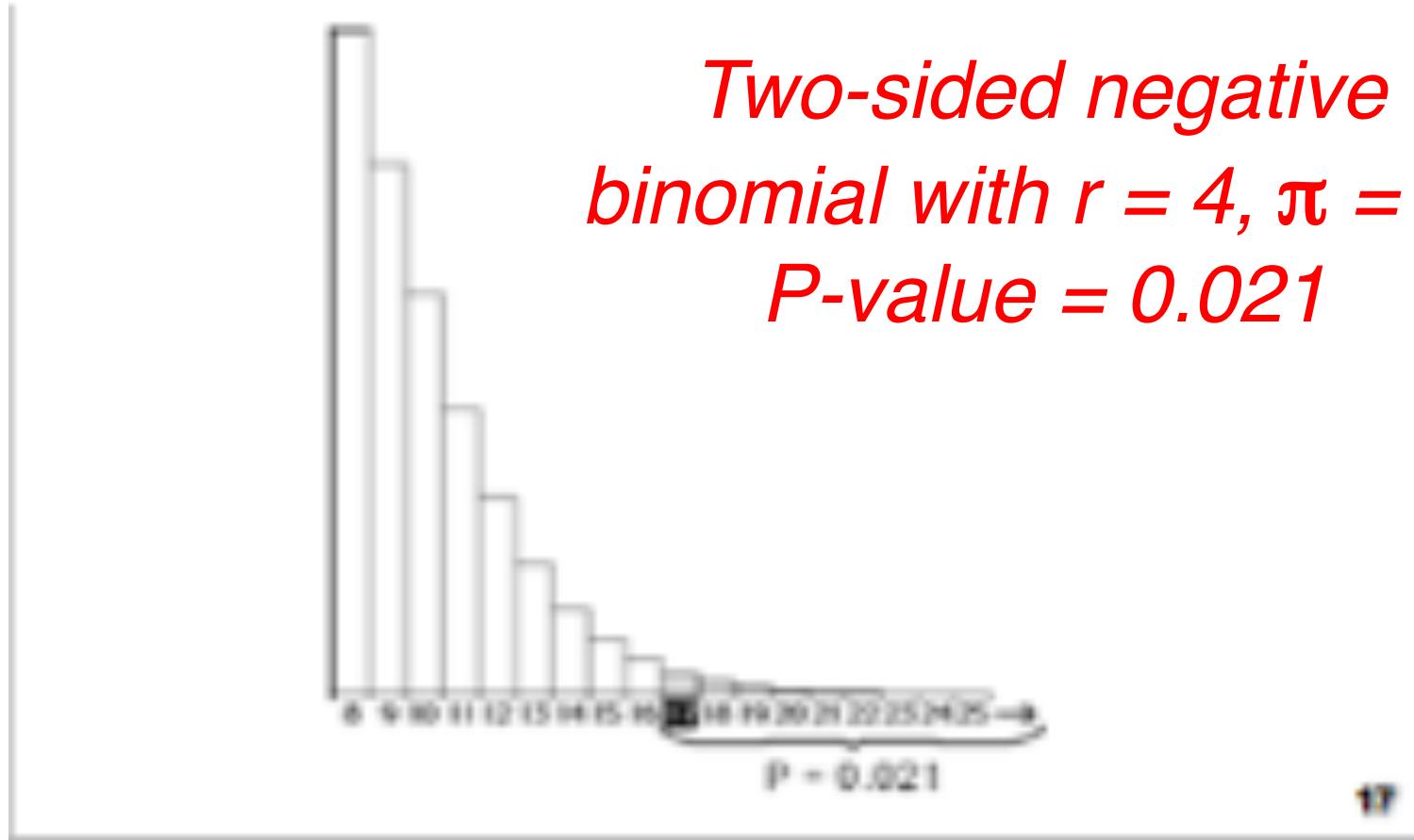
In each design – observe 13/17 successes

Design (1): 17 results

*Binomial distribution
with $n = 17$, $\pi = 0.5$;
 $P\text{-value} = 0.049$*

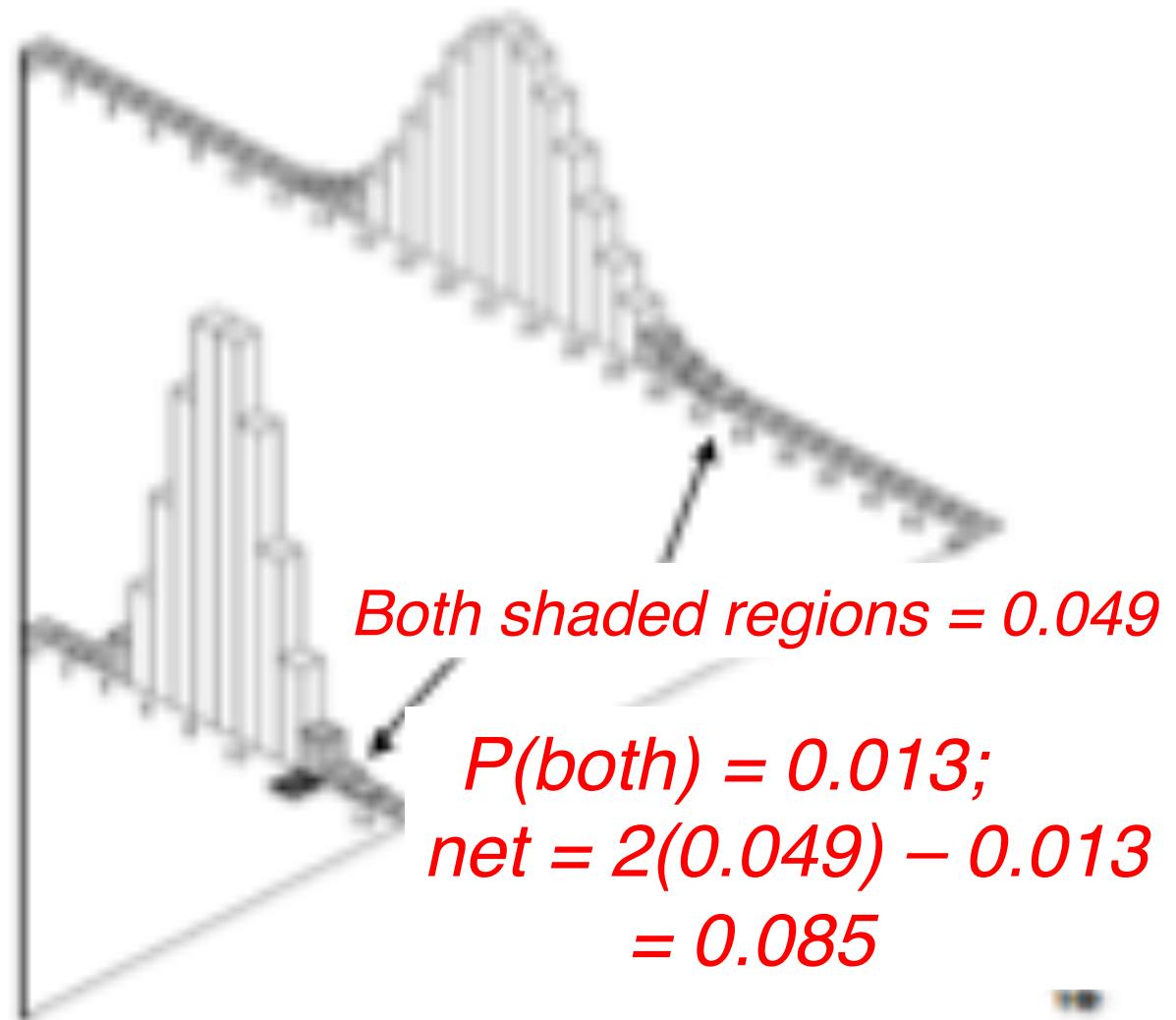


Design (2): Stop when both 4 A's and 4 B's



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

*Analyses at
 $n = 17 \& 44;$
stop @ 17
if 0-4 or 13-17;
 $P = 0.085$*



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

- Cannot calculate P-value
- Frequentist inferences are impossible

Bayesian Stopping Rule

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

Ramifications

- The ramifications to the design of a trial are enormous
 - Almost nobody does freq correctly (too damn hard, silly)
- We have a myth that .. Oh we should only look 1-2 times in a trial! Why?
- Adaptive designs; driven by freq statistics are very very difficult and quickly lose meaning; stopping rules that don't occur—matter to our statistical summary

My House Selection?

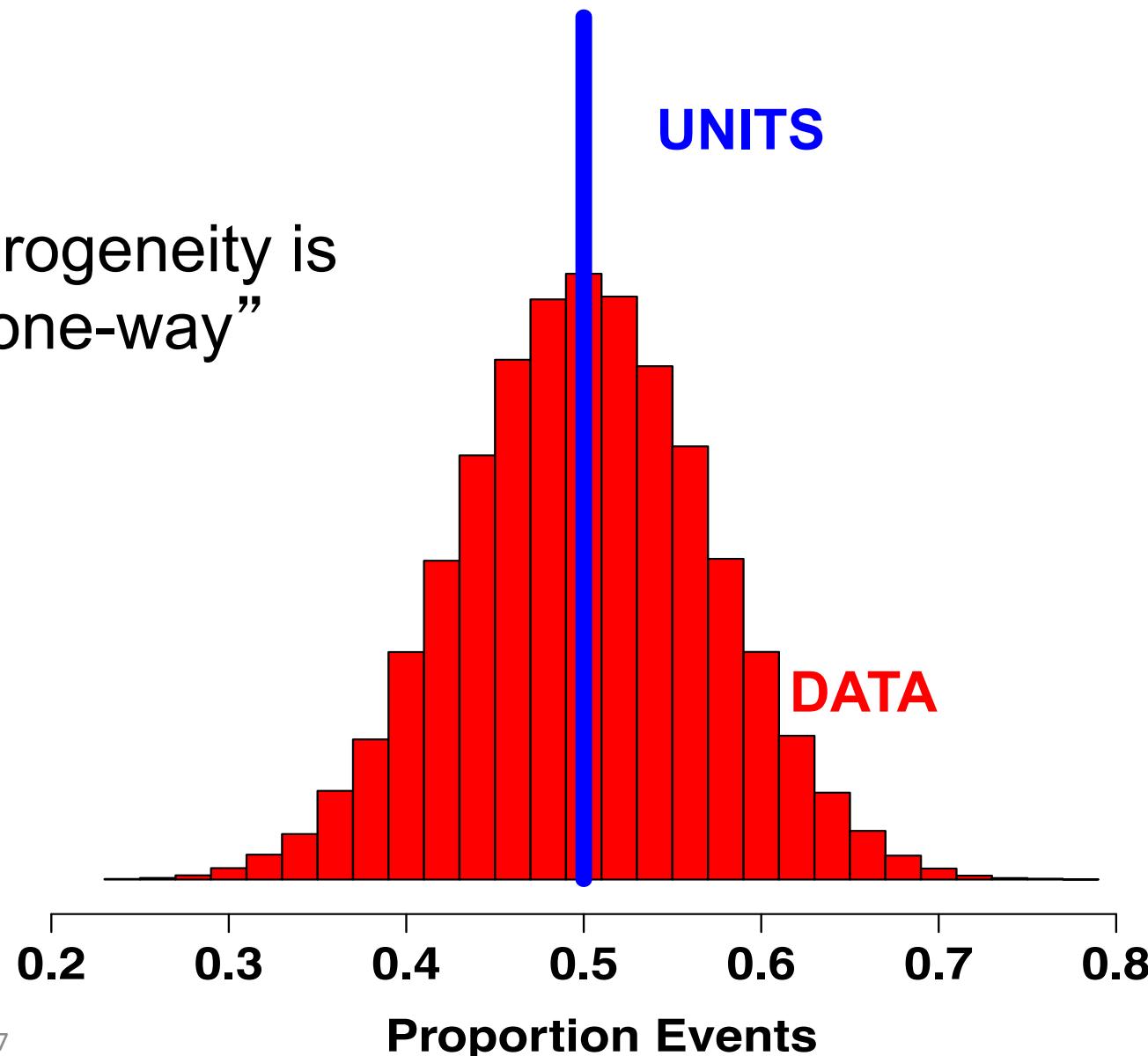


What is Heterogeneity?

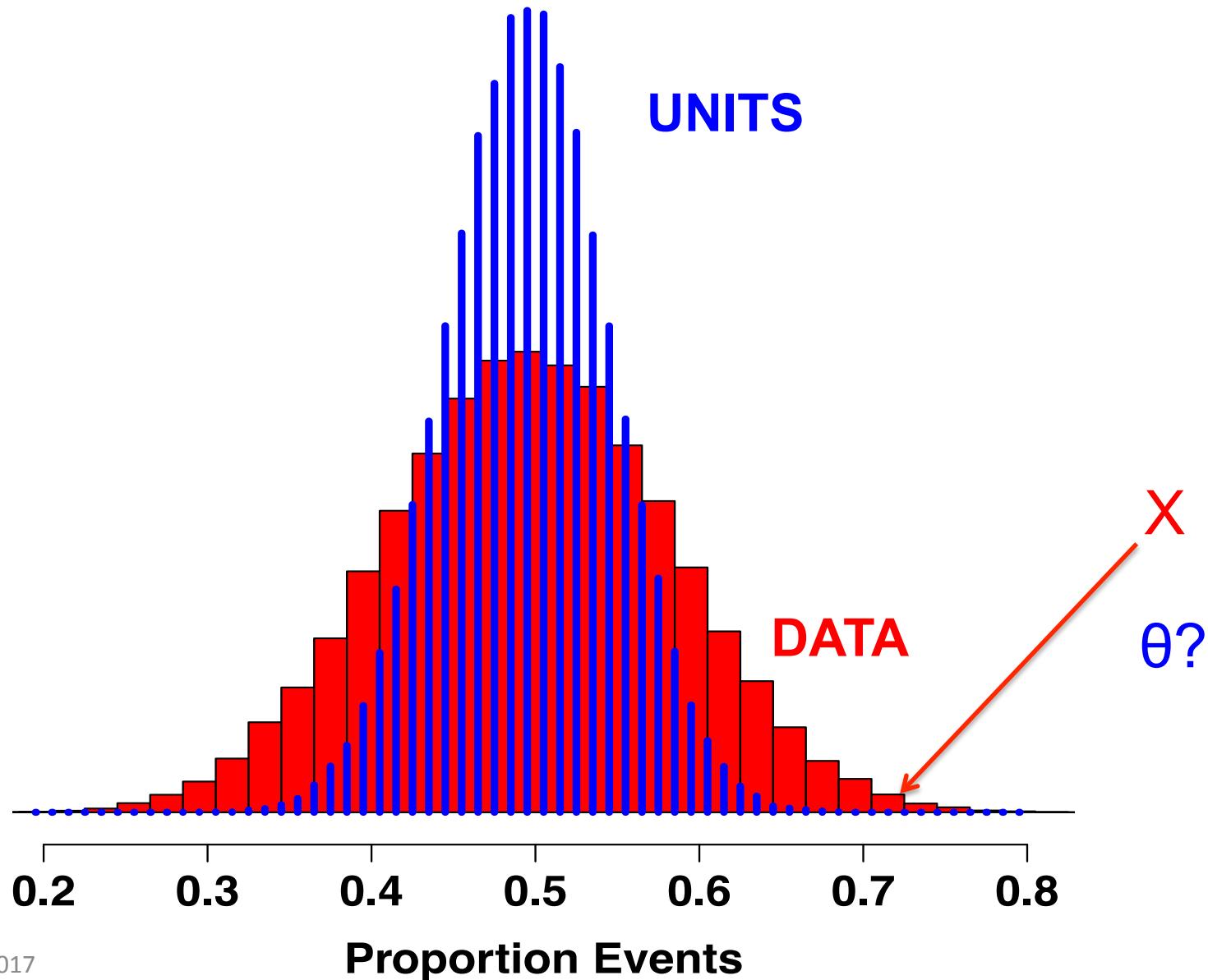
- Webster: *Diverse in character or content*
- Statistically: Variability across units—not data
 - Studies, centers, regions, surgeons, doses, drugs, disease subtypes, demographics, time, indications, ..., *safety outcomes*, ...

Histogram of Data 1

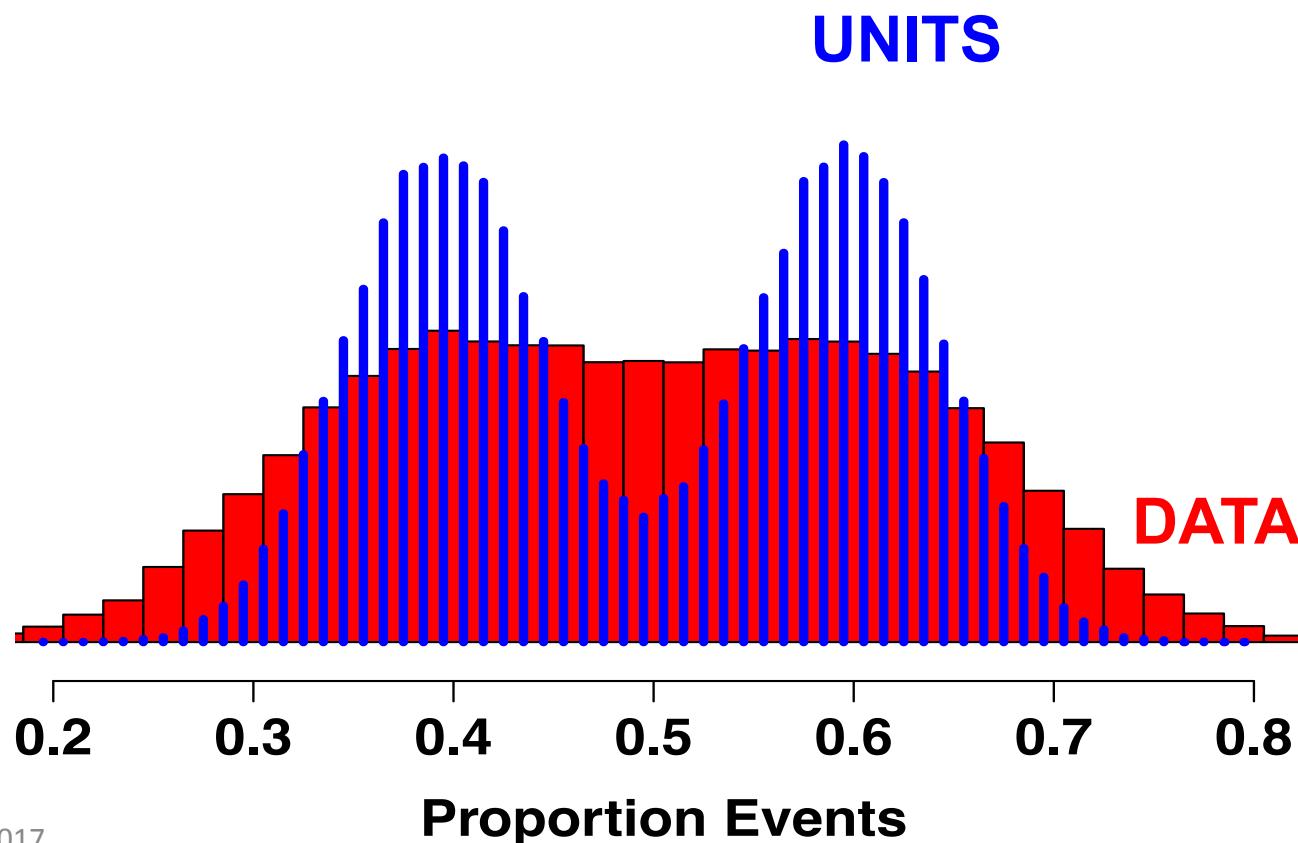
Heterogeneity is
“one-way”



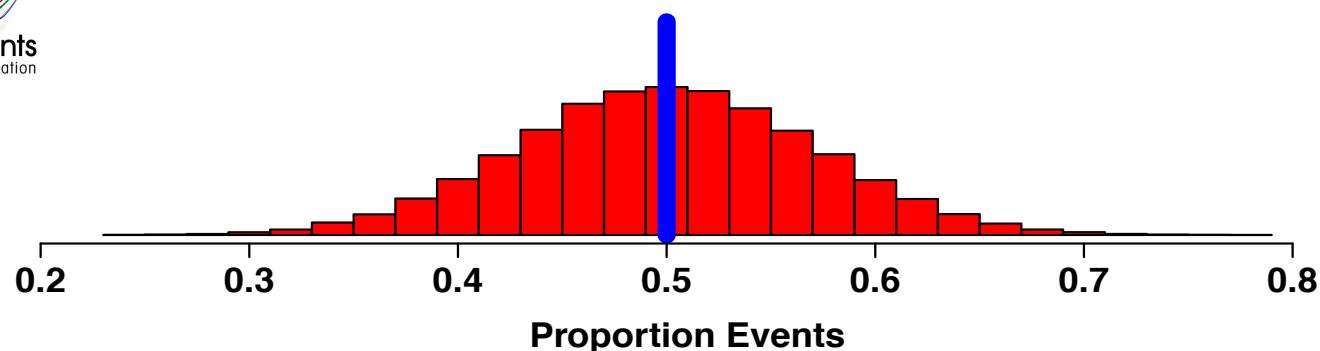
Histogram of Data 2



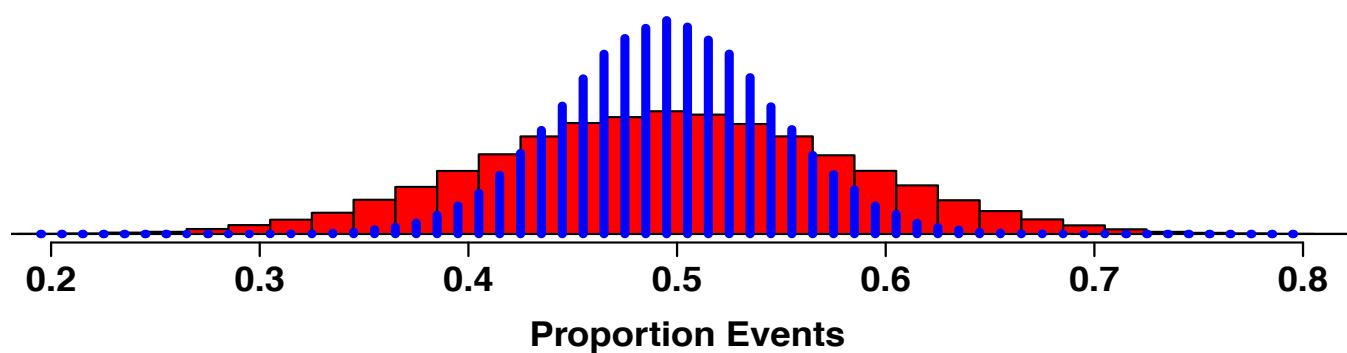
Histogram of Data 3



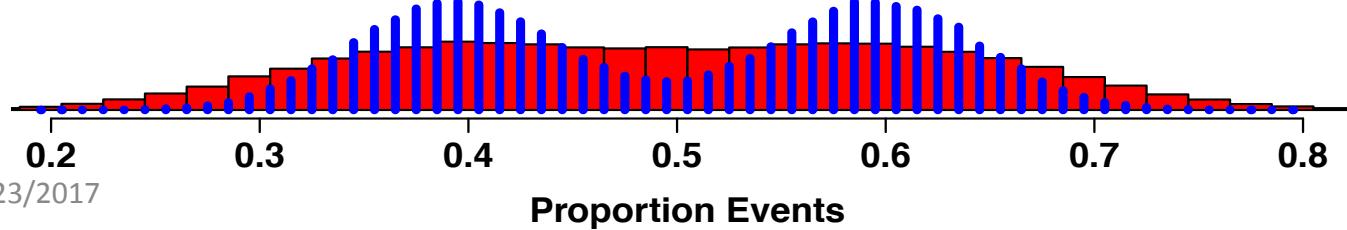
Histogram of Data 1



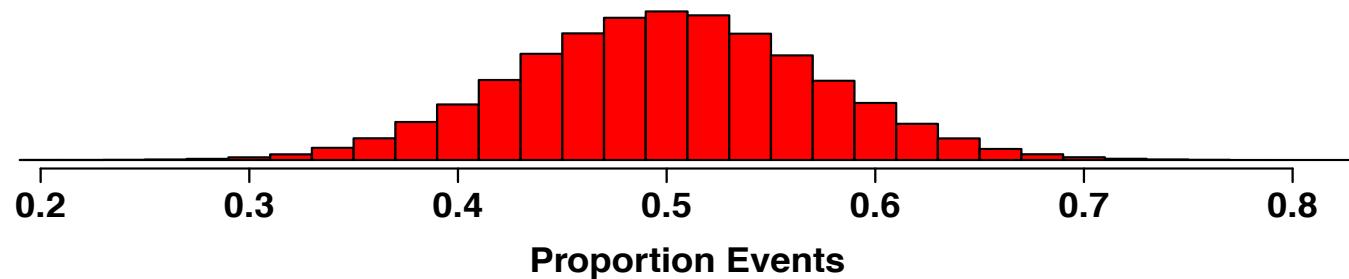
Histogram of Data 2



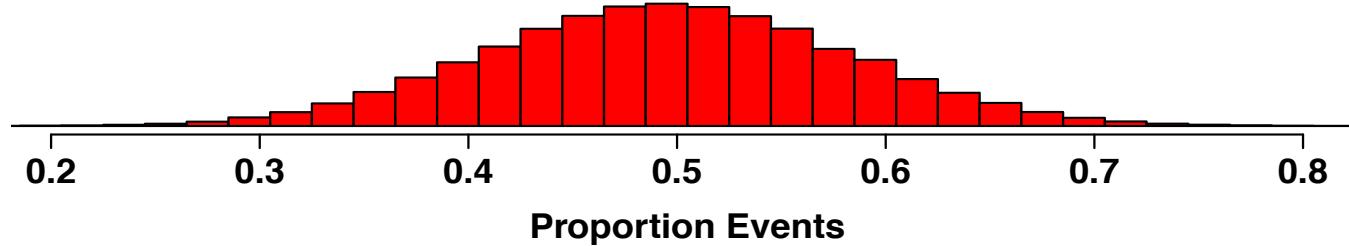
Histogram of Data 3



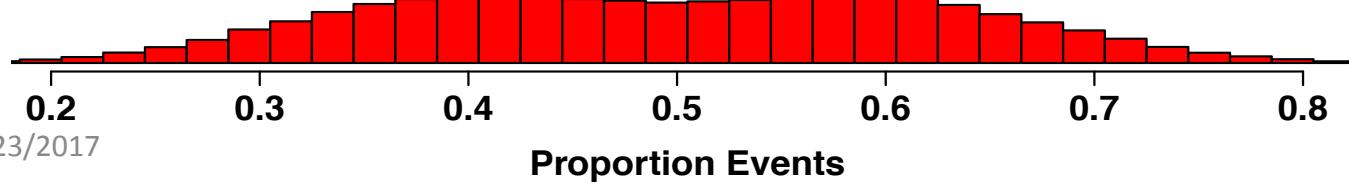
Histogram of Data 1



Histogram of Data 2

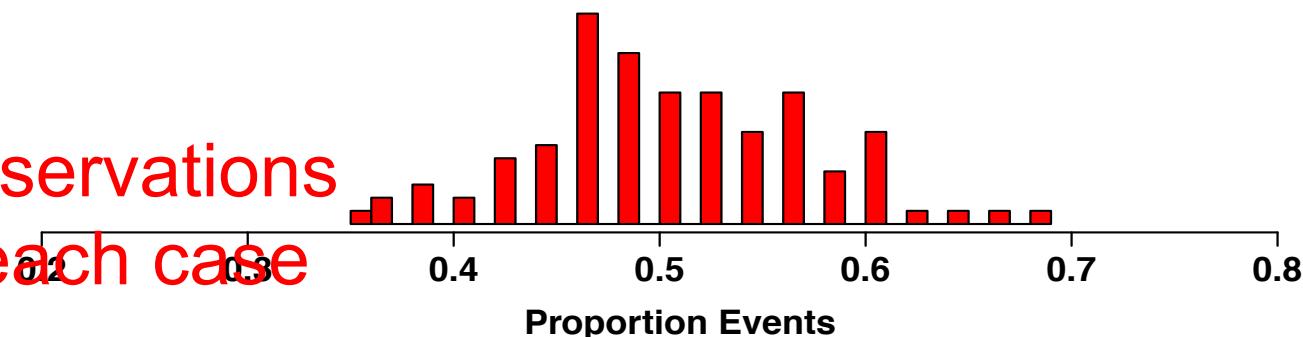


Histogram of Data 3

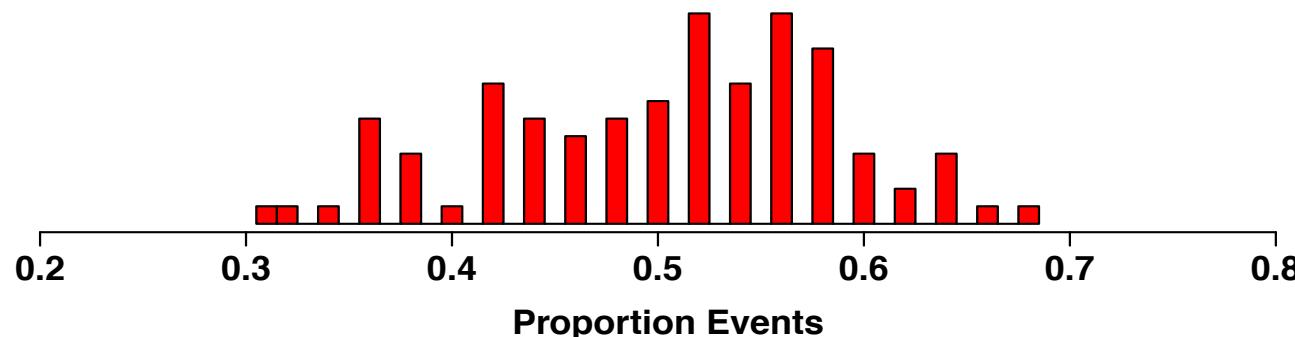


100 observations
from each case

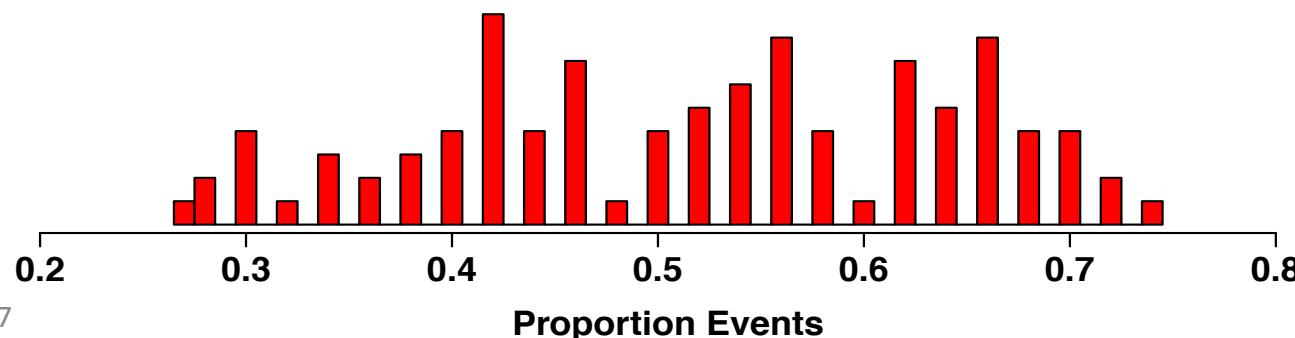
Histogram of Data 1



Histogram of Data 2



Histogram of Data 3



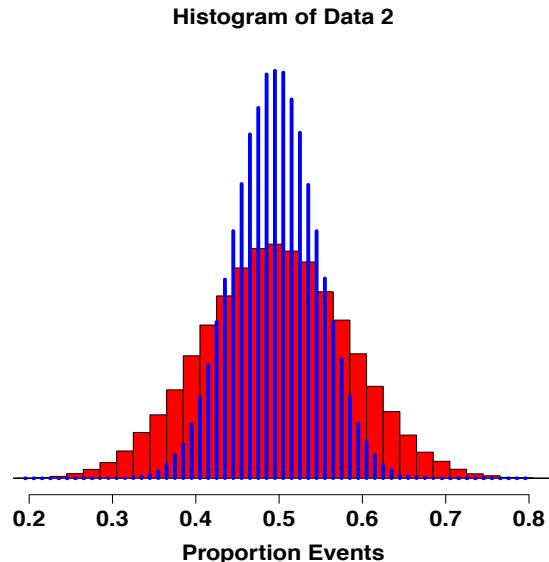
Modeling Heterogeneity

- As a scientist we want to understand/ model heterogeneity
- In some cases it is *the* important question, many times it is a nuisance—It can *drive* the questions
- Differentiate different levels of heterogeneity—levels of units
- What is the *unit of measure*?

Modeling Heterogeneity

- Bayesians naturally model levels of units and variability— inherently synthetic
- Can handle heterogeneity as the question or as a nuisance
- Hierarchical modeling (random effects)
- Ubiquitous in Meta-analysis

Hierarchical Modeling



$$X_i \sim Bin(n_i, \theta_i)$$

$$\theta_i \sim \text{Beta}(\alpha, \beta)$$

$$(\alpha, \beta) \sim f$$

- Hierarchical levels of modeling
- Above is “2-levels”, can be more...
- Allows addressing questions about **red**, **blue**, or **green** levels

CapitalOne™ Lecture Series

THE BILLION DOLLAR STATISTICAL CONCEPT:

How a Lack of Understanding of a 'Simple' Statistical Technique Causes Huge Losses and Poor Decisions in Sports (and Drug Development!)

Scott Berry
President, Berry Consultants

Monday, March 19, 2012
Duncan Hall 1070

ABSTRACT

This talk will discuss hierarchical modeling in sports and present numerous examples when this simple idea is ignored—and the mistakes that are made because of this. Examples from baseball, golf, and many other sports will be explored to show the lack of understanding of a relatively simple statistical idea. This lack of understanding has enormous negative effects in decisions that are made. Interestingly, these decisions are the same as poor decisions that are made in drug development—again resulting in poor decisions—and billions in losses. As more and more “information” is being collected in every field of sport, business, and science, this simple idea is becoming more critical to understand and utilize.

ABOUT THE SPEAKER

Scott Berry is President and Senior Statistical Scientist at Berry Consultants. Since 2000 he has been involved in the design of innovative and adaptive clinical trials in drugs, devices, and biologics. His research interests are in Bayesian methods, computations, simulations, and hierarchical models. In addition to his innovative work in biostatistics he is a renowned sports statistician, with more than 40 articles, ranging from the Journal of the American Statistical Association to ESPN Magazine. He received his PhD from Carnegie Mellon University (1994) and his BS from the University of Minnesota (1990). He spent 5 years at Texas A&M University (1995–2000) before founding Berry Consultants with Don Berry.

A reception in Martel Hall will follow the talk.

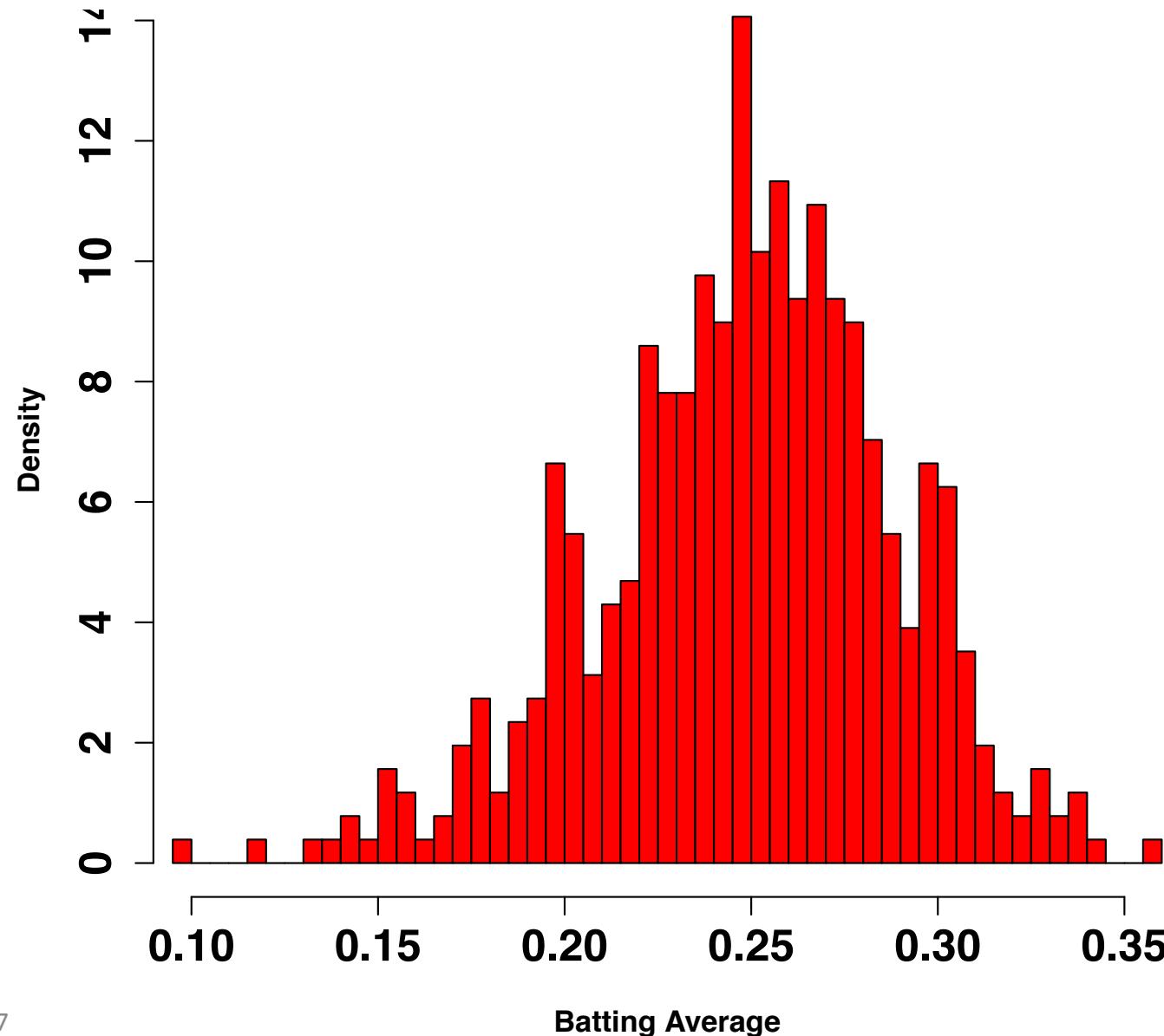


RICE UNIVERSITY
100 YEARS
1912 - 2012

Sponsored by the Department of Statistics and the Center for Computational Finance and Economic Systems



MLB 2011 Batting Averages



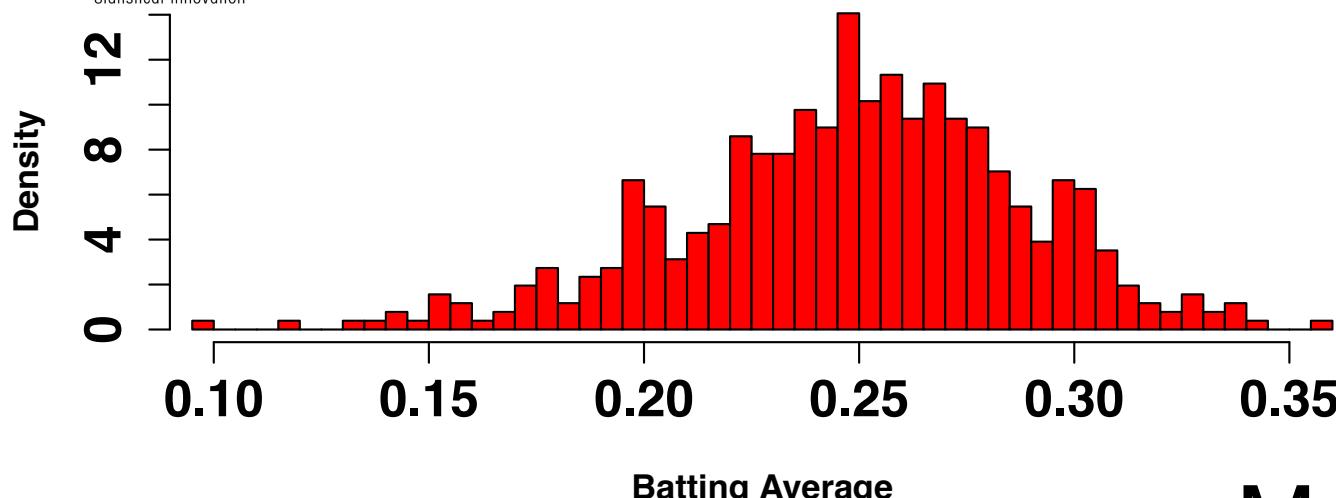
Model

$$X_i \sim Bin(n_i, \theta_i)$$

$$\theta_i \sim \text{Beta}(\alpha, \beta)$$

$$(\alpha, \beta) \sim \exp(-\alpha / 10 - \beta / 10)$$

2011 50+ AB

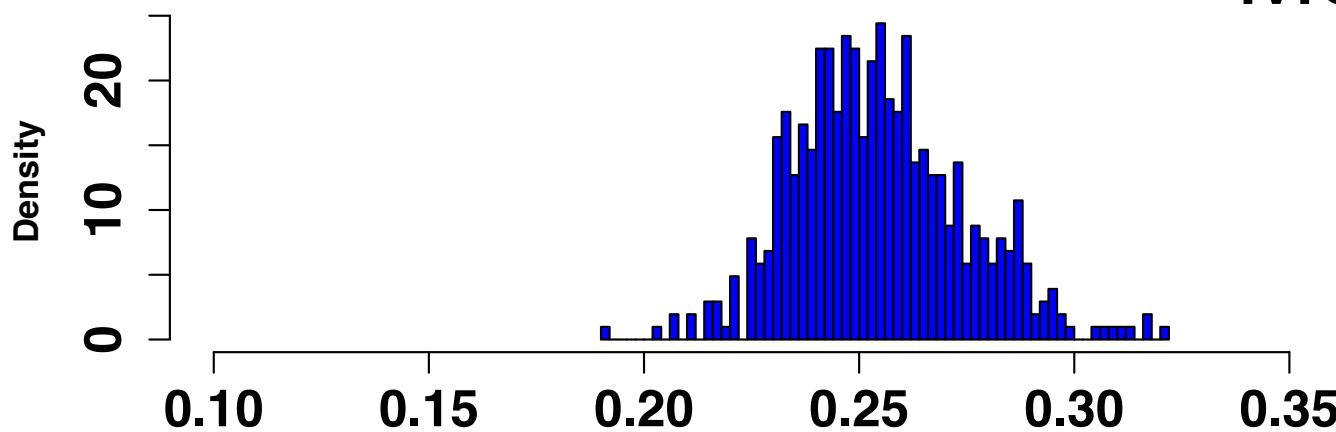


Batting Average

Mean $\alpha = 53.9$

HM Estimates

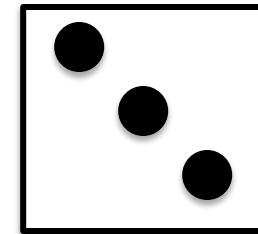
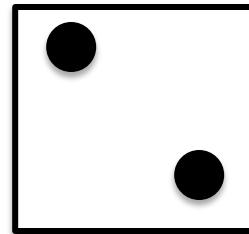
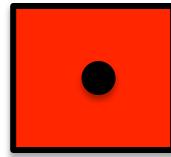
Mean $\beta = 157.8$



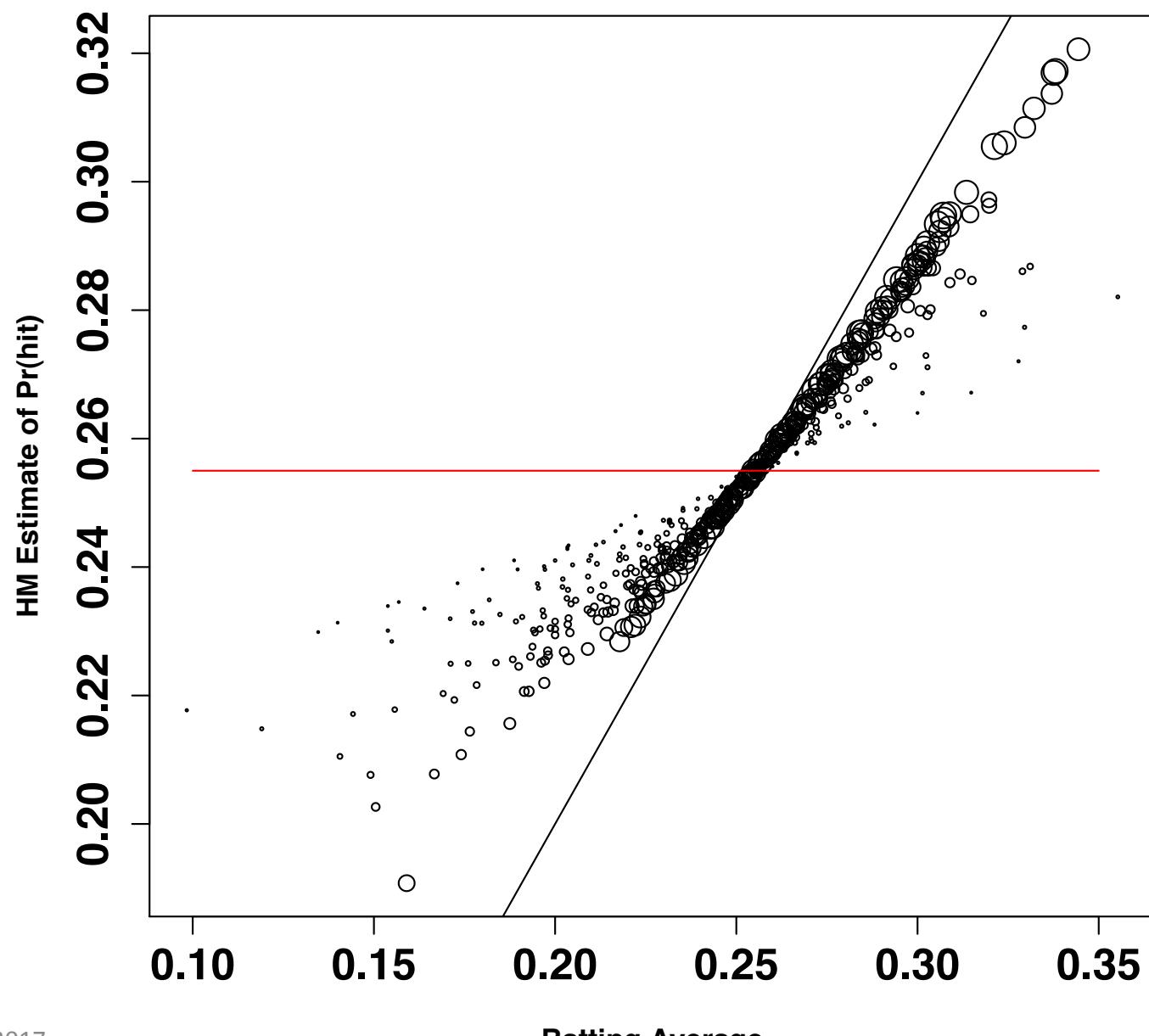
Estimated Pr(Hit)

Strat-O-Matic Baseball

R MARK MCGWIRE 1987			stealing-(E) 4,5/12 (11-4)			bunting-D	hit & run-D
						running	1-10
31% VS.LEFTY PITCHERS - Power-N			69% VS.RIGHTY PITCHERS - Power-N				
1	2	3	1	2	3		
2-lo max	2-fly(lf)B?	>2-gb(ss)B	2-lo max	2-fly(lf)B?	2-gb(ss)A		
+ injury	3-HBP	3-gb(ss)A	+ injury	3-fly(rf)B?	3-gb(ss)B		
3-WALK	4-gb(ss)A+	4-gb(ss)A	3-strikeout	4-gb(ss)A	>4-gb(ss)B		
4-fly(cf)B?	5-strikeout	5-WALK	4-fly(cf)B?	5-strikeout	5-DO**	1-1	
5-HOMERUN	6-SI*	1-3	5-WALK	6-HOMERUN	6-SI*	1-6	SI** 2-20
6-HOMERUN	lineout	4-20	7-WALK	6-HOMERUN	lineout	7-20	6-WALK
7-HR	1-9	7-strikeout	8-strikeout	7-HR	1-6	7-strikeout	7-strikeout
DO	10-20	8-strikeout	9-WALK	DO	7-20	8-strikeout	8-strikeout
#8-fly(lf)B?	9-strikeout	10-SI*	1-6	#8-fly(lf)B?	9-gb(ss)A+		
9-DO**	1-11	10-gb(ss)A+	lineout	9-TR	1-7	10-gb(ss)A	10-SI* 1-6
SI**	12-20	11-fly(rf)B?	>11-SINGLE(lf)	SI**	8-20	11-gb(ss)B+	lineout 7-20
#10-HR	1-15	>12-foulout(c)	\$12-SINGLE(cf)	#10-HR	1-15	12-foulout(c)	>11-SINGLE(cf)
fly(lf)B	16-20			fly(lf)B	16-20		\$12-SINGLE(cf)
11-strikeout				11-HBP			
>12-fly(cf)A				12-fly(cf)A			



HM Mean vs. Observed Estimate



Bayesian Modeling

- Internal company prediction
 - Meta-analysis; forecasts; PK/PD forecasts
- Investor forecasts
 - Stunning benefit to HM
- Disease Progression Modeling
 - Rare diseases
 - Alzheimer's
 - DIAD (random-effects modeling)
 - “progressive diseases”
- Categorical outcomes/Dose-Response
- Re-Randomization in Oncology

“If we could first know where we are, and whither we are tending, we could then better judge what to do, and how to do it.”

--Abraham Lincoln,
“*House Divided*” speech
June 16, 1858



Bayesian Calculations

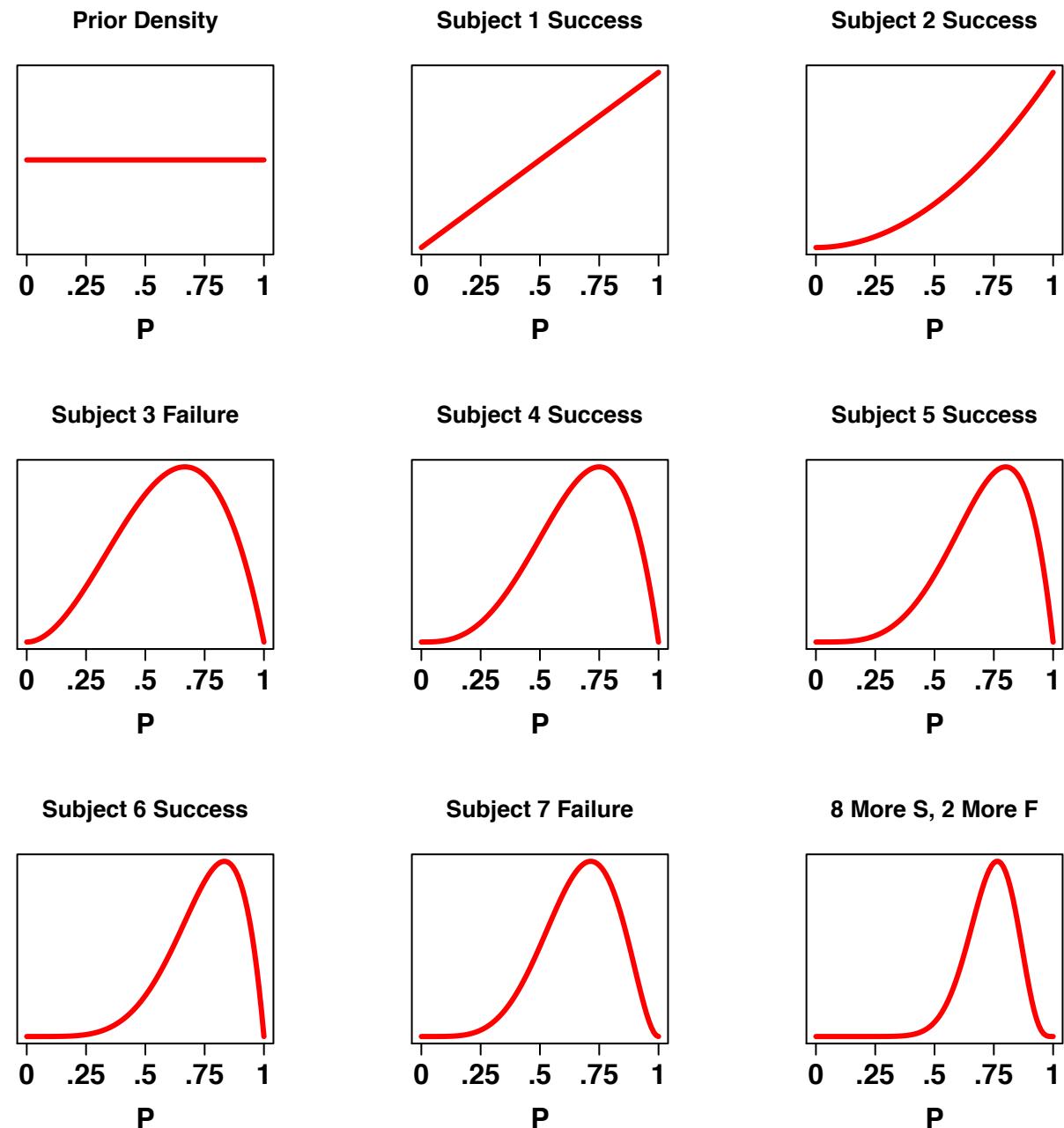
- Data: 13 S's and 4 F's
- For **ANY** design with these results, the likelihood function is

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

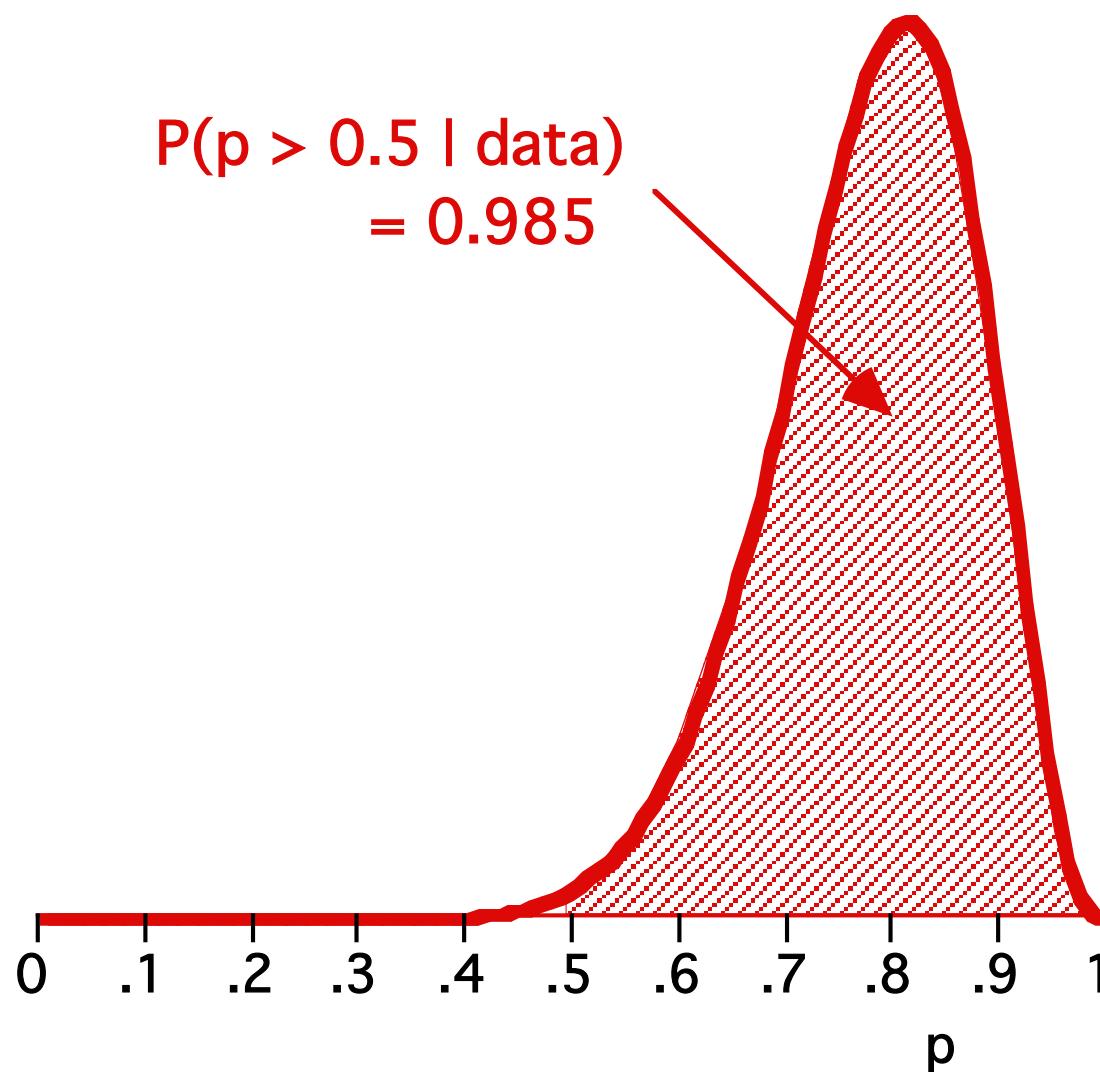
- Posterior probabilities...
 - Let's assume a Beta(1,1)....

Bayesian Analyses of All...

Or
Updated sequentially



$$\Pr[\pi > 0.5]$$



PREDICTIVE PROBABILITIES

- Distribution of future data?
- $P(\text{next is an F}) = ?$
- Critical component of experimental design
- In monitoring trials

Predictive Distribution

- The posterior distribution of a future observation of $X_i\dots$

$$[x_{n+1} | x_1, \dots, x_n] = \int [x_{n+1} | \theta][\theta | x_1, \dots, x_n] d\theta$$

- 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

Suppose 17 more observations

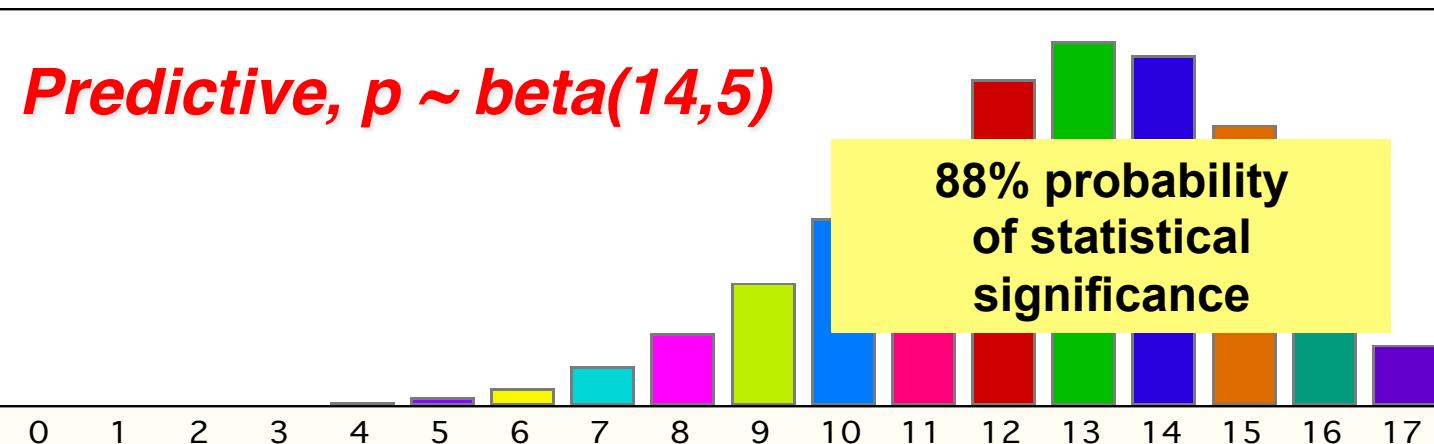
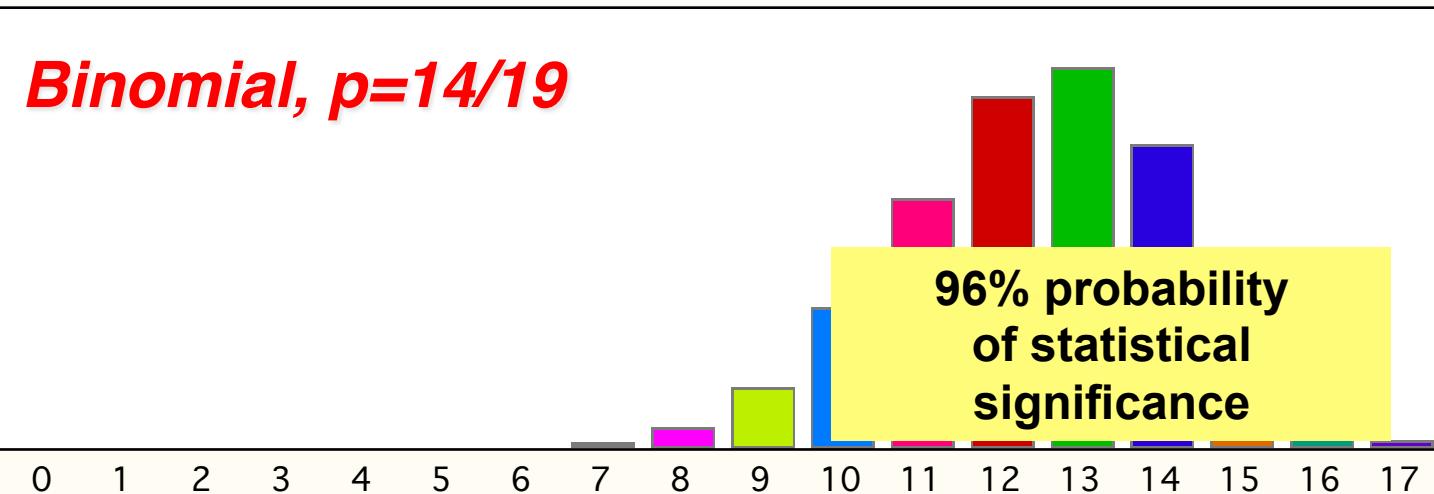
$P(S \text{ on } x \text{ of next 17} / \text{data})$

(23/34 wins trial)

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

Beta-Binomial Distribution

Best fitting binomial vs. predictive probabilities



Possible Calculation

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

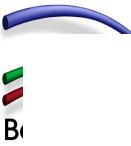
- Simulate a π from the $\text{beta}(14,5)$
- Simulate an x from $\text{binomial}(17, \pi)$
- Distribution of x 's is beta-binomial--the predictive distribution

Posterior and Predictive...same?

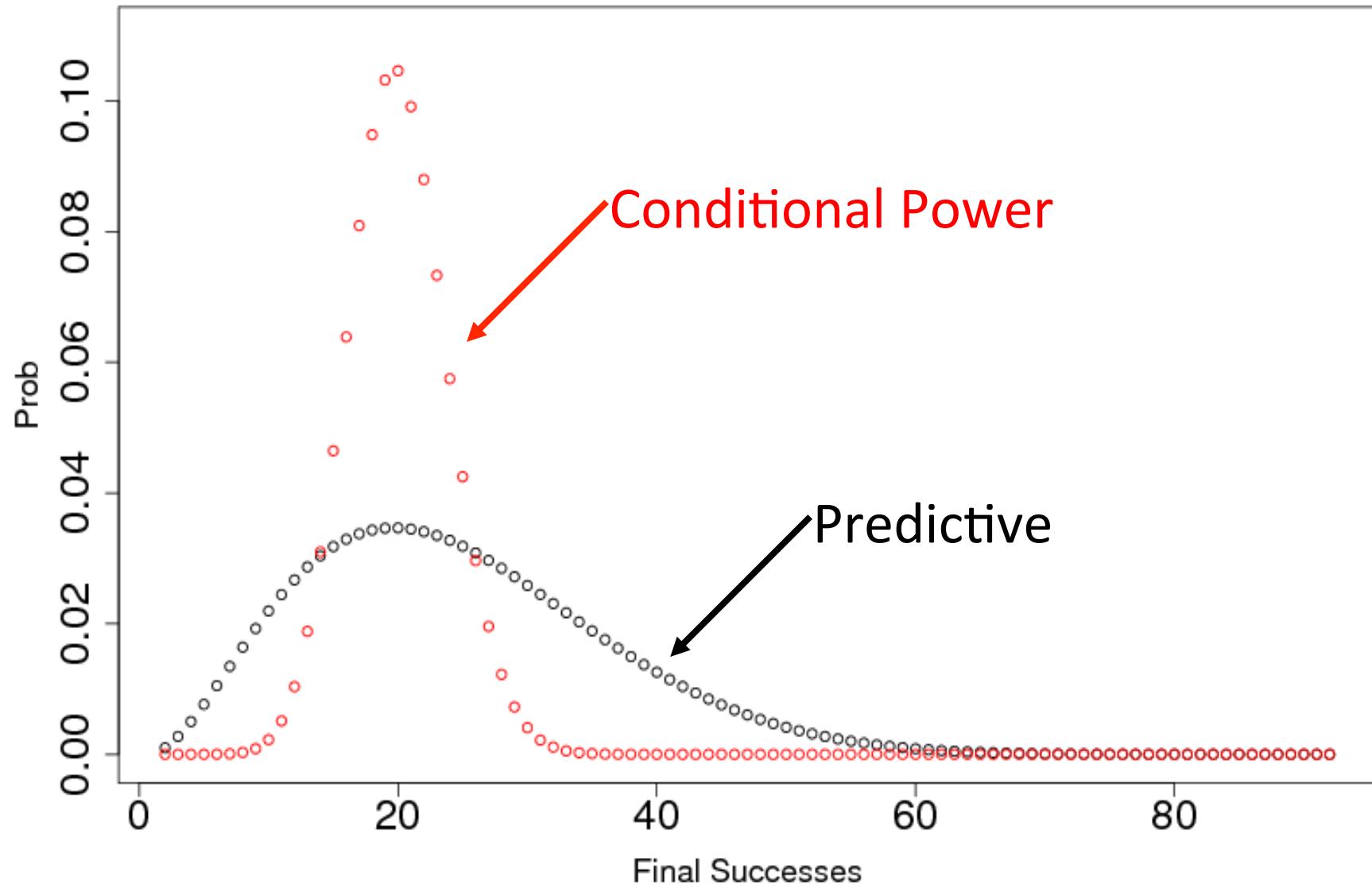
- Clinical Trial, 100 subjects. $H_A: \pi > 0.25$? FDA will approve if # success ≥ 33 [post > 0.95 , beta(1,1)]
- @ interim, 99 subjects, 32 successes
- $\Pr[\pi > 0.25 | \text{data}] = 0.955$
- Predictive prob trial success = 0.327

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



Predictive Distr'n if 2/10

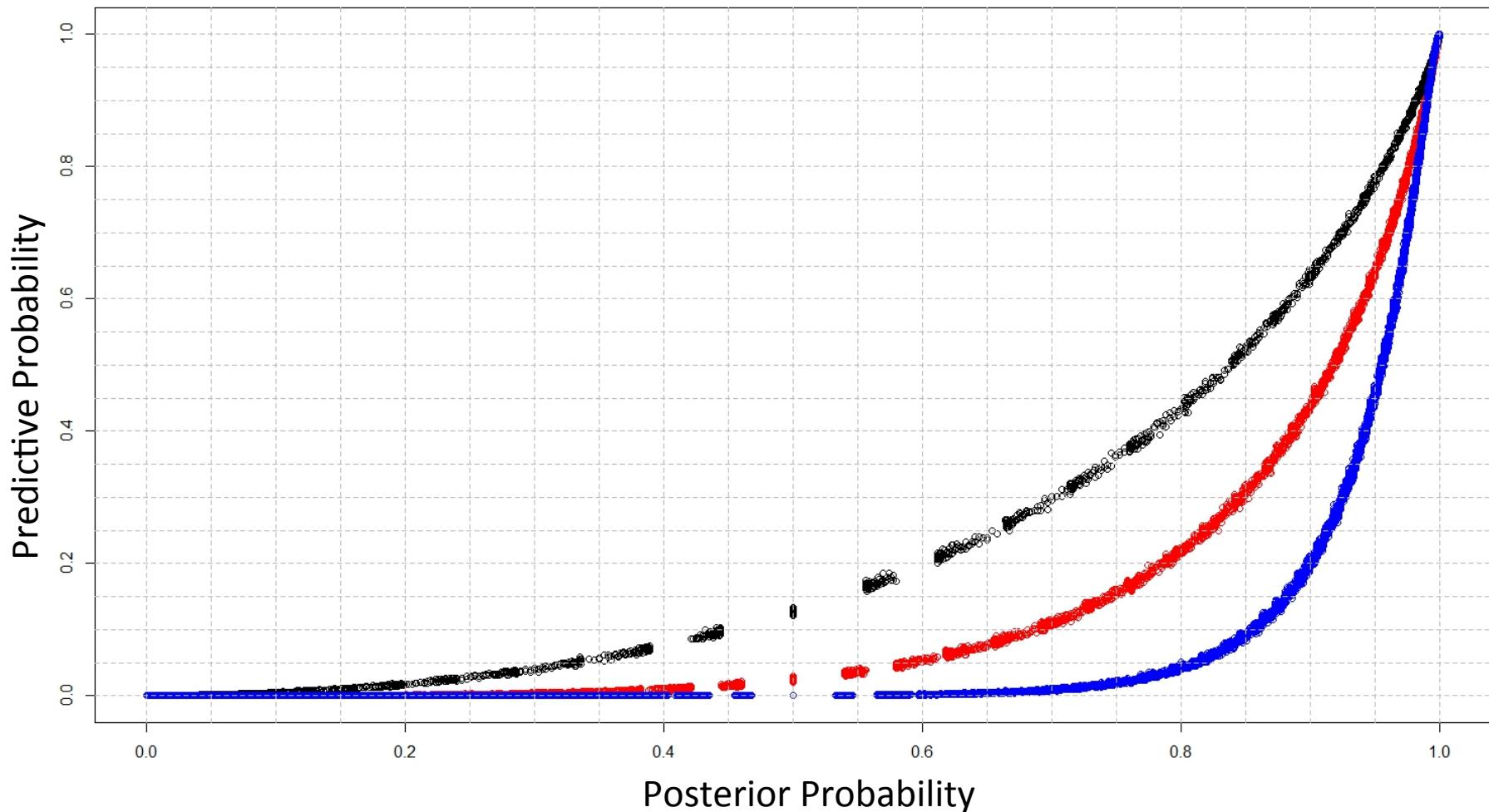


Predictive, Posterior, CP

S@10	Post Prob >0.25	PP	CP
0	.042	.0096	0
1	.197	.070	6.6×10^{-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

Posterior versus (Current) Predictive

for n=100 (black), 200 (red), 300 (blue) per arm



“Borrowed” from Kert Viele, Anna McGlothlin, Liz Krackey

DIA:6/23/2017

52

Interpretation

- Predictive is VERY different than posterior probability
- If you were using frequentist CP 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”

VALOR, An Adaptive Design, Pivotal Phase 3 Trial Of Vosaroxin Or Placebo In Combination With Cytarabine In First Relapsed Or Refractory Acute Myeloid Leukemia



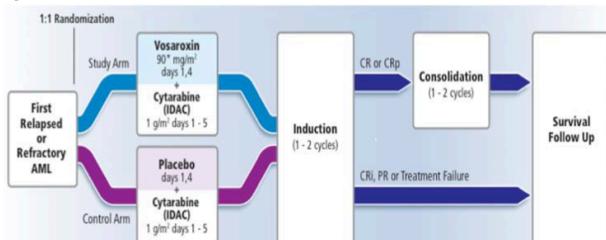
Farhad Ravandi, MD¹; Ellen K. Ritchie, MD²; S. Hamid Sayar, MD³; Stephen A. Strickland, MD⁴; Michael D. Craig, MD⁵; Dominik Selleslag, MD⁶; Johan Maertens, MD⁷; Violaine Havelange, MD⁸; Jeffrey Lancet, MD⁹; Gary Acton, MD¹⁰; Cyrus Mehta, PhD¹¹; Eric Silva¹²; Adam Craig, MD, PhD¹⁰; Judith A. Fox, PhD¹⁰; Robert K. Stuart, MD¹³; Harry P. Erba, MD, PhD¹⁴; Norbert Vey, MD¹⁵; Gary J. Schiller, MD¹⁶; Eric J. Feldman, MD¹⁷

¹Univ of Texas MD Anderson Cancer Ctr, Houston, TX; ²Cornell Medical Center, New York, NY; ³Indiana University Cancer Center, Indianapolis, IN; ⁴Vanderbilt-Ingram Cancer Center, Nashville, TN; ⁵West Virginia University, Morgantown, WV; ⁶AZ St. Jan Brugge, Brugge, Belgium; ⁷UZ Gasthuisberg, Leuven, Belgium; ⁸Cliniques Universitaires Saint Luc, Brussels, Belgium; ⁹Moffitt Cancer Center, University of South Florida, Tampa, FL; ¹⁰Sunesis Pharmaceuticals, Inc., South San Francisco, CA; ¹¹Cytel, Inc., Cambridge, MA; ¹²Medical Univ of South Carolina, Charleston, SC; ¹³Division of Hematology and Oncology, Univ of Michigan, Ann Arbor, MI; ¹⁴Institut Paoli-Calmettes, Marseille, France; ¹⁵Univ of California, Los Angeles, CA

VALOR TRIAL DESIGN

VALOR (NCT01191801), a pivotal phase 3, randomized, controlled, double-blind trial, evaluates vosaroxin and cytarabine versus placebo and cytarabine in patients with first relapsed or refractory acute myeloid leukemia (AML) incorporating an adaptive design. The primary endpoint is overall survival (OS); secondary/tertiary endpoints include complete remission (CR) rates, safety, event free survival (EFS), leukemia free survival (LFS), and transplantation (HSCT) rate.

Figure 1. VALOR Trial Schema



* After cycle 1, all subsequent cycles at 70 mg/m² vosaroxin on days 1 and 4

Sample Size	450 evaluable patients
Population	First relapsed or refractory AML
Regimen	IDAC + vosaroxin vs. IDAC + placebo (double-blind)
Study Sites	>110 sites in Europe, North America, AUS/NZ
Interim Analysis	Single, pre-planned evaluation by DSMB
Adaptive Design	At interim analysis, DSMB can recommend adding 225 evaluable patients to the trial

Key Eligibility Criteria

- At least 18 years old with an AML diagnosis by WHO classification
- First relapsed AML with first CR or CRp (CR1) duration of at least 90 days to 24 months OR refractory AML with persistent leukemia after 1 or 2 induction cycles or CR1 less than 90 days
- No more than 2 prior induction cycles that include at least 1 regimen of cytarabine with an anthracycline (or anthrancenedione)
- Adequate cardiac, hepatic and renal function
- Refractory to or relapsed within the previous 3 months after therapy with an IDAC- or HIDAC-containing regimen

VALOR ADAPTIVE DESIGN

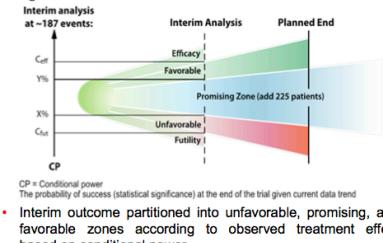
	Base Case:	Alternative Case:
Power	90% power to detect a 40% survival difference (5 vs. 7 mo.)	90% power to detect a 30% survival difference (5 vs. 6.5 mo.)
Hazard ratio and α	0.71 and 0.05 (2-sided)	0.77 and 0.05 (2-sided)
Resources needed	375 OS events from 450 evaluable patients	562 OS events from 675 evaluable patients
Enrollment	24 months with 6 months follow-up	30 months with 6 months follow-up

• Vosaroxin + cytarabine arm Base Case treatment effect is supported by phase 2 in first relapsed or primary refractory AML(N=69)

- Median OS 6.9 mo.
- Combined CR rate 29% (CR rate 26%)
- Median LFS (defined as time from CR to relapse or death) 24 mo.
- 30 and 60 day all-cause mortality 3% and 9%, respectively
- HSCT rate 26%
- Control arm median OS of 5 mo. is based on published IDAC-based regimens outcomes
- Alternative Case provided a scenario with smaller but meaningful treatment effect
 - Other scenarios adequately powered under VALOR adaptive study design

DSMB RECOMMENDATIONS BASED ON INTERIM RESULT

Figure 3. Outcomes Based on Conditional Power at Interim



CP = Conditional power
The probability of success (statistical significance) at the end of the trial given current data trend

- Interim outcome partitioned into unfavorable, promising, and favorable zones according to observed treatment effect based on conditional power

VALOR RECOVERS POWER BY SAMPLE SIZE INCREASE IF IN PROMISING ZONE

True Hazard Ratio	OS Improvement	Base Case Design 450 Patients, 375 Events	Adaptive Design 675 Patients, 562 Events
0.71	40%	91%	98%
0.74	33%	83%	96%
0.77	30%	71%	90%
0.80	25%	58%	84%

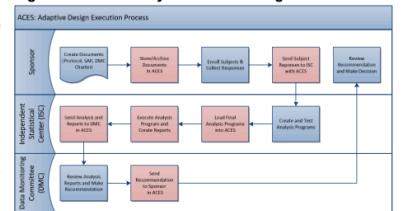
• VALOR's adaptive design gains substantial additional power over non-adaptive IF interim outcome falls in the Promising Zone

PROTECTING INTEGRITY OF ADAPTIVE DESIGN TRIAL

- Guidance documents by FDA and EMA for DSB and Adaptive Trial Design:
 - Reference the importance of confidentiality of interim results
 - Suggest "A well-trusted firewall established for trial conduct can help provide assurance that statistical and operational biases have not been introduced."
 - Requests an accurate recording of trial conduct and documentation – who saw what and when

ACCESS CONTROL EXECUTION SYSTEM (ACES)

Figure 4. Interim Analysis Process Using ACES



VALOR STATUS AND SUMMARY

- VALOR IS enrolling well with 317 patients as of May 14, 2012
 - On track to conduct pre-specified interim analysis in Q3 2012
 - DSMB recommended VALOR continue as planned after reviewing safety data in Dec 2011
- VALOR is a well-powered study designed to detect a clinically meaningful improvement in OS
 - DSMB may call for sample size increase only if interim result falls into Promising Zone
 - The adaptive design mitigates risk of initial over-investment, and risk of failing to detect a relevant survival benefit
 - This design satisfies both the statistical and operational requirements stipulated in FDA Draft Guidance and in EMA Reflection Paper on Adaptive Design Clinical Trials

VALOR, An Adaptive Design, Pivotal Phase 3 Trial Of Vosaroxin Or Placebo In Combination With Cytarabine In First Relapsed Or Refractory Acute Myeloid Leukemia



Farhad Ravandi, MD¹; Ellen K. Ritchie, MD²; S. Hamid Sayar, MD³; Stephen A. Strickland, MD⁴; Michael D. Craig, MD⁵; Dominik Selleslag, MD⁶; Johan Maertens, MD⁷; Violaine Havelange, MD⁸; Jeffrey Lancet, MD⁹; Gary Acton, MD¹⁰; Cyrus Mehta, PhD¹¹; Eric Silva¹²; Adam Craig, MD, PhD¹⁰; Judith A. Fox, PhD¹⁰; Robert K. Stuart, MD¹³; Harry P. Erba, MD, PhD¹⁴; Norbert Vey, MD¹⁵; Gary J. Schiller, MD¹⁶; Eric J. Feldman, MD²

¹Univ of Texas MD Anderson Cancer Ctr, Houston, TX; ²Cornell Medical Center, New York, NY; ³Indiana University Cancer Center, Indianapolis, IN; ⁴Vanderbilt-Ingram Cancer Center, Nashville, TN; ⁵West Virginia University, Morgantown, WV; ⁶AZ St. Jan Brugge, Brugge, Belgium; ⁷UZ Gasthuisberg, Leuven, Belgium; ⁸Cliniques Universitaires Saint Luc, Brussels, Belgium; ⁹Moffitt Cancer Center, University of South Florida, Tampa, FL; ¹⁰Sunesis Pharmaceuticals, Inc., South San Francisco, CA; ¹¹Cytel, Inc., Cambridge, MA; ¹²Medical Univ of South Carolina, Charleston, SC; ¹³Division of Hematology and Oncology, Univ of Michigan, Ann Arbor, MI; ¹⁴Institut Paoli-Calmettes, Marseille, France; ¹⁵Univ of California, Los Angeles, CA

VALOR TRIAL DESIGN

VALOR ADAPTIVE DESIGN

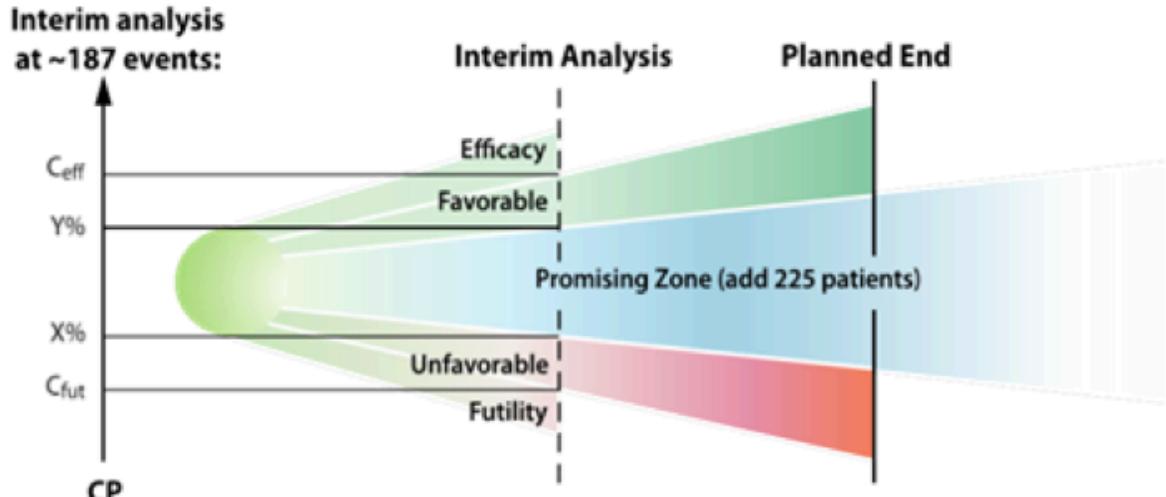
VALOR RECOVERS POWER BY SAMPLE SIZE INCREASE IF IN PROMISING ZONE

DSMB RECOMMENDATIONS BASED ON INTERIM RESULT

DSMB can recommend based on interim results to:

- Continue the trial to 450 evaluable patients (375 events)
- Adjust sample size to 675 evaluable patients (562 events)
- Stop early for efficacy ($p<0.0015$) or futility

Figure 3. Outcomes Based on Conditional Power at Interim



CP = Conditional power

The probability of success (statistical significance) at the end of the trial given current data trend

- Interim outcome partitioned into unfavorable, promising, and favorable zones according to observed treatment effect based on conditional power

Interim

- The public announcement was made that at the interim the result was “Promising Zone”
- What does that mean for predicting success?
- Can we calculate predictive probability of success based on learning data = ‘PZ’?
- We did this in 2014...
- We only had partial information on what rules were used... bottom of HR for PZ???
- I think we used a slightly too small value...(other info)

Bayes Theorem by Simulation

Draw Prior HR

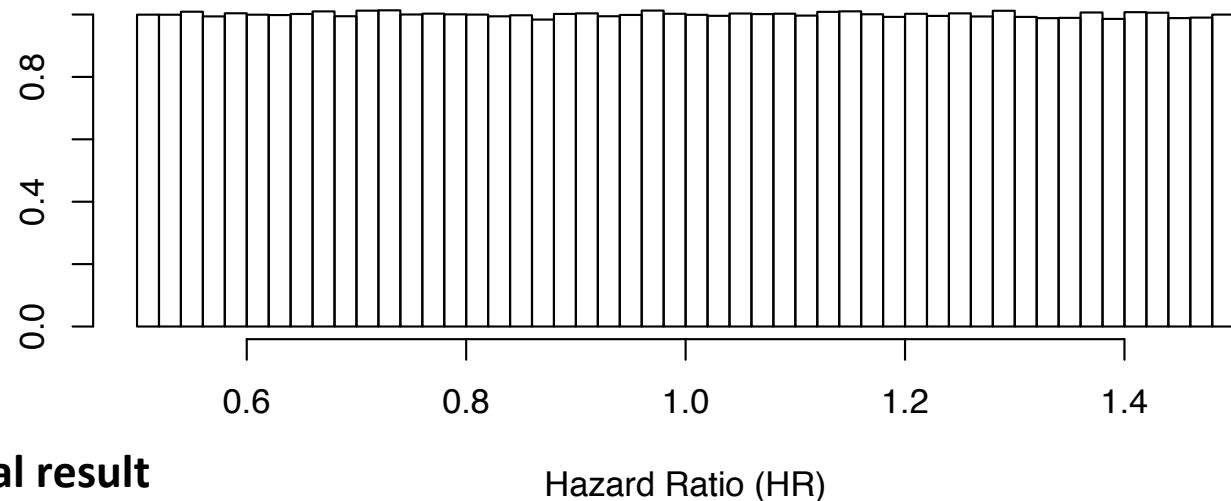
Run trial, see interim & final result

Condition on PZ at interim..

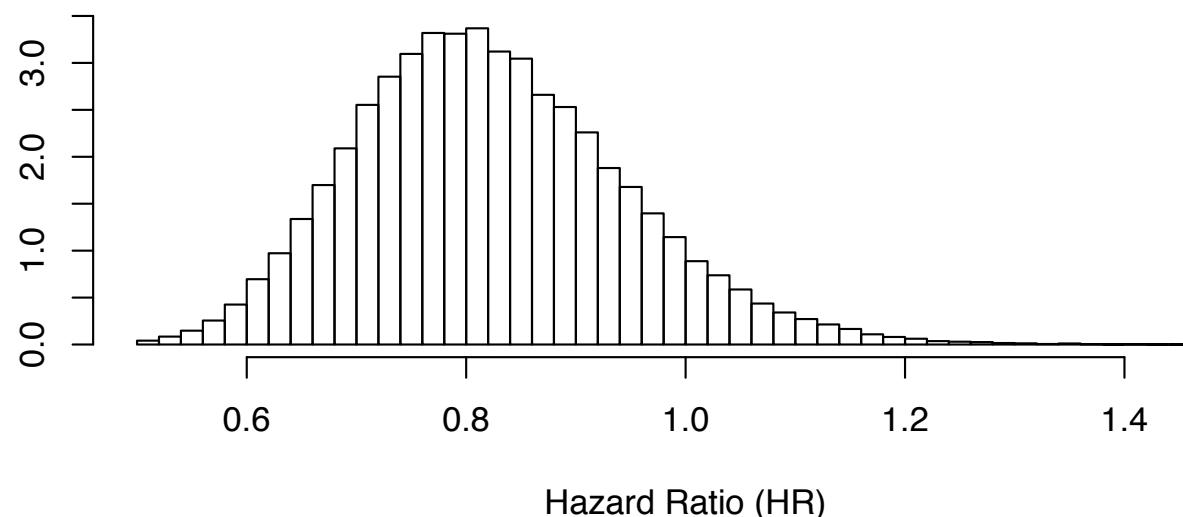
what HR generated that value??

what was the final result?

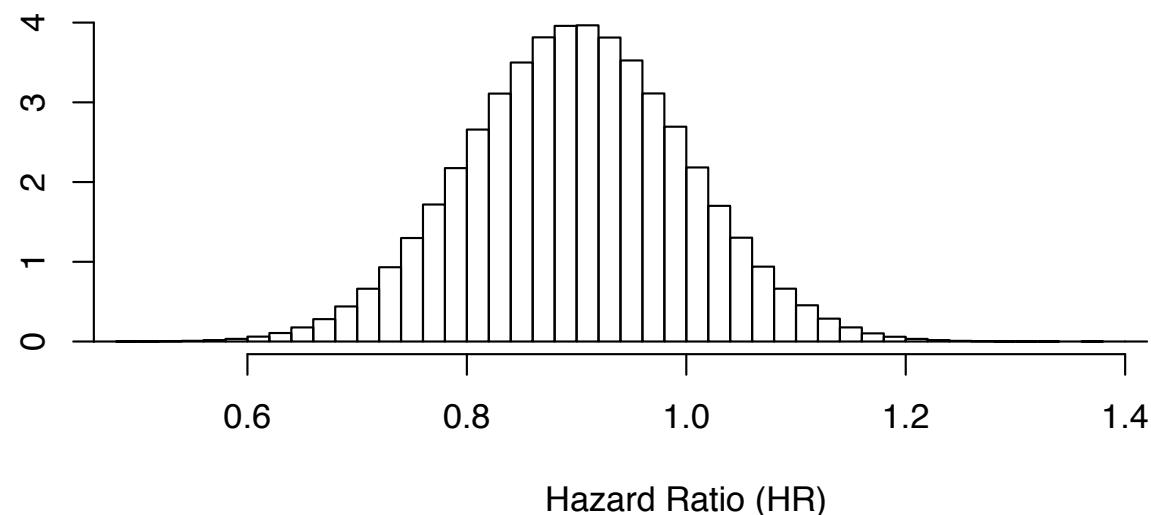
Histogram of Prior for HR: Win= 0.329



Posterior for HR | PZ: Win Prob 0.646

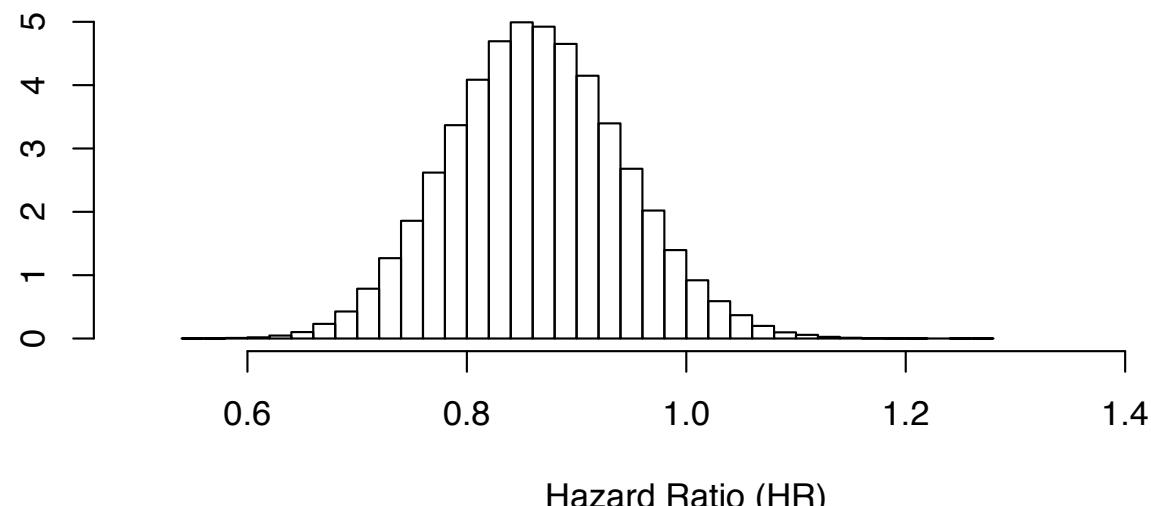


Histogram of Prior for HR: Win= 0.287



This was the “Berry Prior”

Posterior for HR | PZ: Win Prob 0.528



Sunesis Announces Results From Pivotal Phase 3 VALOR Trial of Vosaroxin and Cytarabine in Patients With First Relapsed or Refractory Acute Myeloid Leukemia

October 6, 2014 6:31 AM ET

Trial Does Not Reach Primary Endpoint of Statistically Significant Improvement in Overall Survival

Shows Significant Survival Benefit when Censored for Transplant

Safety Profile Consistent with that Observed in Previous Company Trials

Company Plans to Commence European Filing and Explore U.S. Regulatory Pathway

Sunesis to Host Conference Call and Webcast Today at 8:30 AM Eastern Time

SOUTH SAN FRANCISCO, Calif., Oct. 6, 2014 (GLOBE NEWSWIRE) -- Sunesis Pharmaceuticals, Inc. (Nasdaq:SNSS) today announced results from the pivotal Phase 3 VALOR trial, a randomized, double-blind, placebo-controlled trial of vosaroxin and cytarabine in patients with first relapsed or refractory acute myeloid leukemia (AML). At more than 100 leading international sites, the trial enrolled 711 patients, who were stratified for age, geography and disease status. The trial did not meet its primary endpoint of demonstrating a statistically significant improvement in overall survival, with a median overall survival of 7.5 months for vosaroxin and cytarabine compared to 6.1 months for placebo and cytarabine (HR=0.865, p=0.06). Because transplant may confound the primary analysis, a predefined analysis of overall survival censoring for stem cell transplantation was planned. In this analysis, patients receiving the vosaroxin combination had a median overall survival of 6.7 months versus 5.3 months for placebo and cytarabine (HR=0.809, p=0.02). The trial also demonstrated a clinically significant benefit in complete remission (CR) rate (30.1% vs 16.3%, p=0.0000148), the secondary endpoint.

(HR=0.865, p=0.06).

The “Clash of Philosophies”

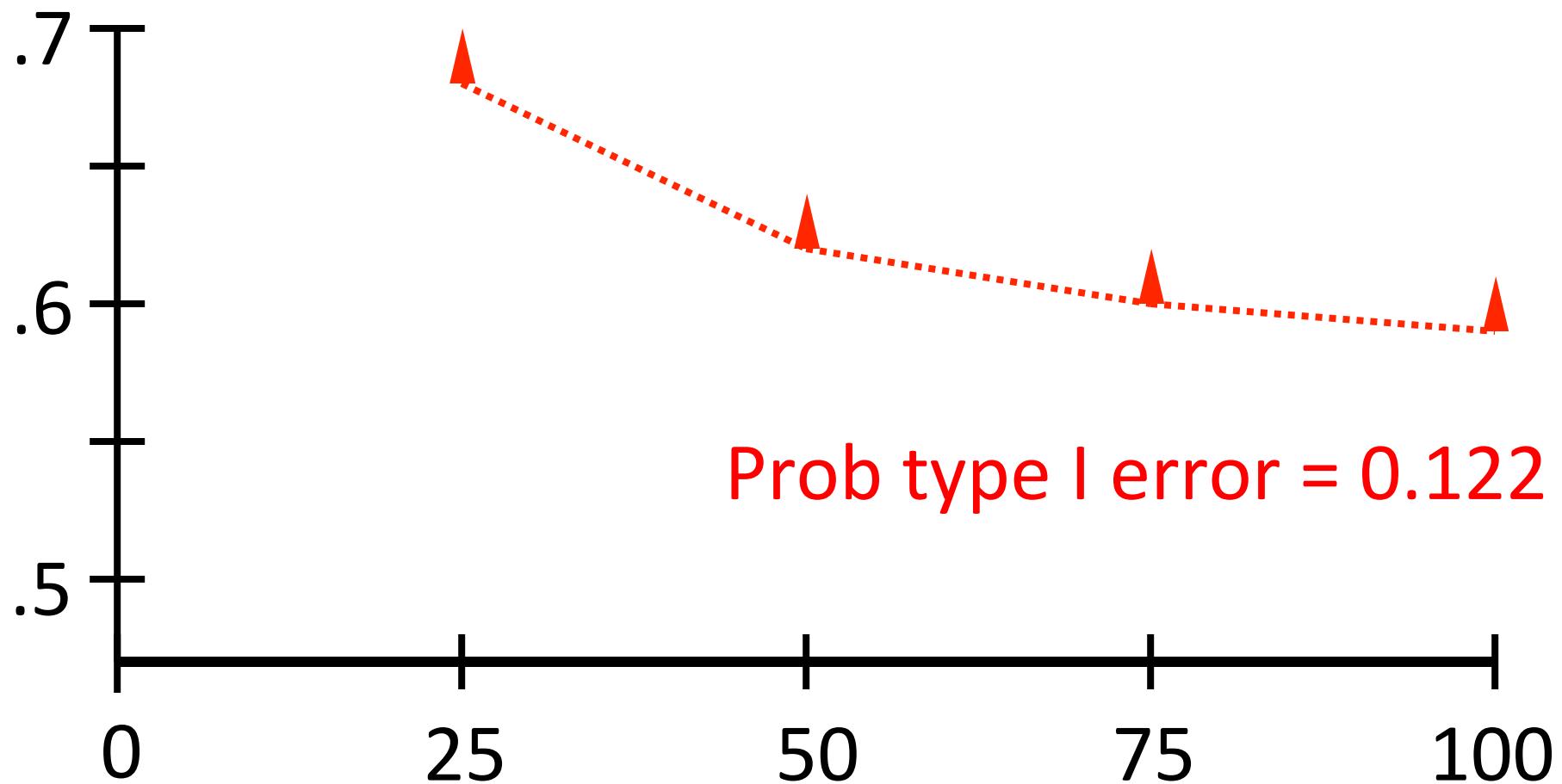
- Single-arm trial, n=100 maximum trial size.
- $X \sim \text{Binomial}(n, p)$
- If $\Pr[p > 0.5 \mid \text{data}] \geq 0.95$ then trial success

$$p \sim \text{Beta}(1, 1)$$

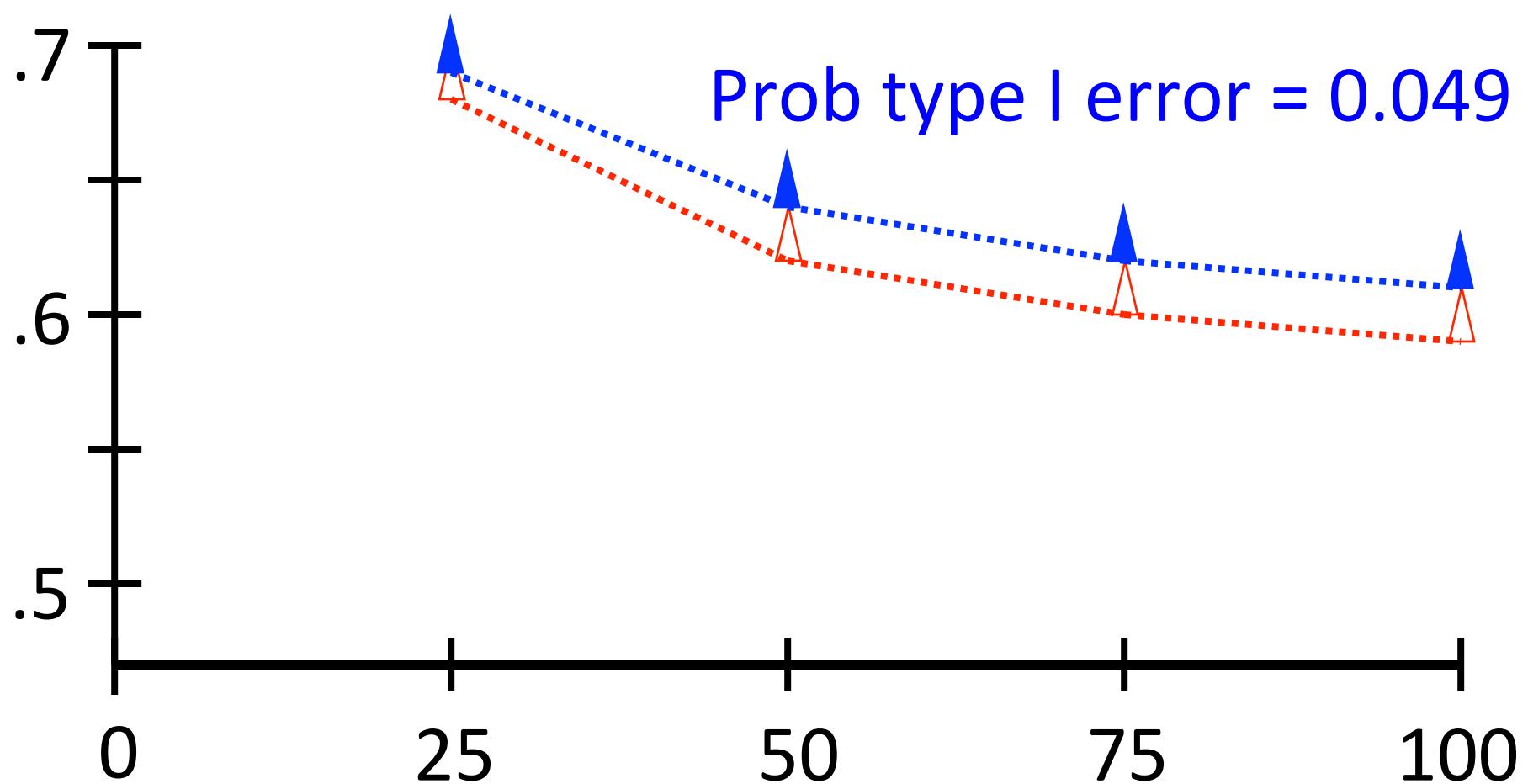
- At n=100, if there are **59** or more successes then trial success
- Type I error is *calculated*: $\sum_{i=59}^{100} \binom{100}{i} \left(\frac{1}{2}\right)^{100} = 0.044$

- Add in looks at n=25, 50, 75.

If $\Pr[p > 0.5] \geq 0.95$ at **any look** then stop for success (17,31,45,59)



If $\Pr[\pi > 0.5] \geq 0.98$ at any look then stop
for success (18,33,47,61)



Simple Example With Priors

- Same scenario as before, but say, we have “pure” information and prior of Beta(10,2) (9/10 data) instead of Beta(1,1).
- *Regulatory agrees it is reasonable to use this as the prior*
- Fixed design: for $\Pr[p > 0.5 \mid \text{data}] \geq 0.95$ then trial success $X \geq 55$ (old 59)
- Type I error of “new” experiment is 18.4% (“combined” is 5.23%)

The Result

- Solution: Raise the post prob bar to 0.9934 to get Type I error below 0.05!!!
- Need a Beta(59+10,41+2) for a win...**59 is back!!!**
- The type I error “restriction” forces 59/100 regardless of prior...
- **Can't allow positive priors AND force type I of “new” experiment!**
- **Essentially conditioning on anything (unless negative data) creates type I error increase moving forward**

Bayesian

- The root of Bayesian decision making is having a goal, or utility function to maximize with a decision
 - The posterior means nothing by itself...
- Can we select a treatment that maximizes the utility for a patient? Utility for the sponsor?

- Modified Rankin Score (mRS) is frequently used as an endpoint in neurological assessments (stroke, etc)

Example

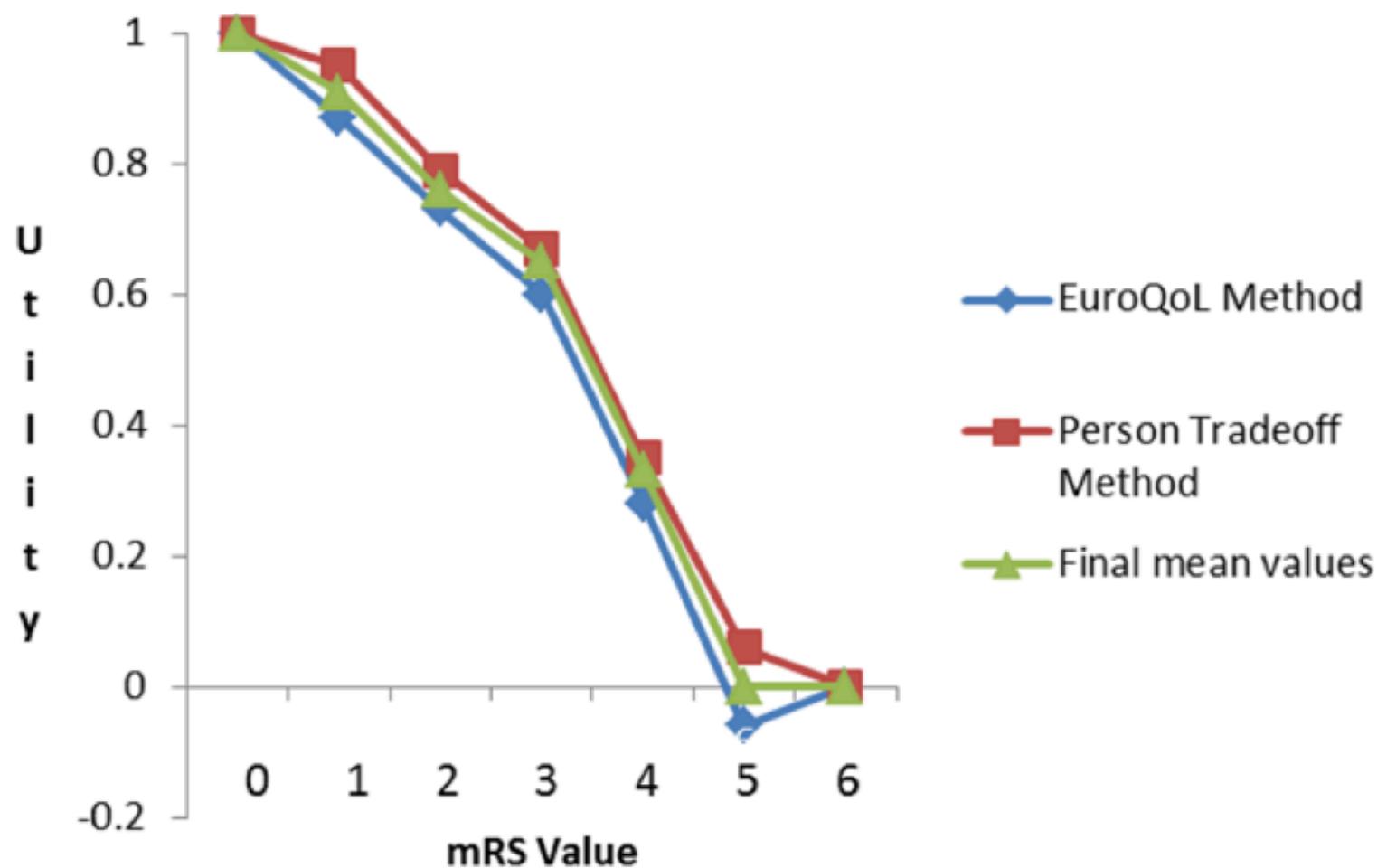
MODIFIED RANKIN SCALE (mRS)		Patient Name: _____
Score	Description	Rater Name: _____
0	No symptoms at all	Date: _____
1	No significant disability despite symptoms; able to carry out all usual duties and activities	
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance	
3	Moderate disability; requiring some help, but able to walk without assistance	
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance	
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention	
6	Dead	
TOTAL (0–6): _____		

mRS

- How to analyze?
 - Dichotomize? Where?
 - Non-parametric?
 - Ordinal proportional model?
 - Resampling approaches
 - Utility of each outcome to the patient?

Utility

mRS	U
0	1
1	0.91
2	0.76
3	0.65
4	0.33
5-6	0



Analysis

- The analysis is then a model for the probability of each outcome (like a Dirichlet over 7 outcomes or Maybe dose-response models?)
- The superiority or analysis is then which treatment has maximum utility; probability one treatment has better utility than another

Effect Direction	Intervention	Trial	Utility weighted mRS (t test)	Ordinal mRS (Mann Whitney)	Dichotomized 0-1 vs 2-6	Dichotomized 0-2 vs 3-6	Dichotomized 0-4 vs 5-6
Unidirectional	IV TPA	NINDS TPA Trials	0.0031	0.001	0.00	0.003	0.276
Unidirectional	IV TPA	IST 3	0.041	0.021	0.056	0.43	0.15
Unidirectional	Endovascular 1 st Generation	PROACT 2	0.44	0.33	0.177	0.047	1.0
Unidirectional	Endovascular 1 st Generation	MELT	0.33	0.19	0.03	0.28	0.89
Unidirectional	Endovascular 2 nd Generation	SWIFT	0.011	0.012	0.308	0.308	0.004
Unidirectional	Endovascular 2 nd Generation	MR CLEAN	0.0006	0.0006	0.037	0.0007	0.08
Unidirectional	BP Lowering	INTERACT 2	0.045	0.044	0.033	0.063	0.201
Unidirectional	Surgery	Hemicraniectomy Meta-Analysis	0.0001	0.00002	1.00	0.039	0.00
<hr/>							
Bidirectional	IV TPA	ECASS 3	0.28	0.059	0.037	0.136	0.535
Bidirectional	Endovascular 2 nd Generation	TREVO 2	0.22	0.35	0.047	0.009	0.535
<hr/>							
Neutral	Endovascular 1 st Generation	IMS 3	0.24	0.67	0.58	0.56	0.19

Chaisinanunkul, et al (2015)

Stryker's DAWN Trial Enrollment Ended After Interim Review Indicates Success

[SHARE](#) | [E-MAIL](#) | [PRINT](#) | [BOOKMARK](#) |

March 9, 2017—Stryker announced an early end to patient enrollment in the [DAWN](#) clinical trial comparing mechanical thrombectomy with the company's Trevo Retriever plus medical therapy versus medical therapy alone when initiated within 6 to 24 hours after time last known well. The study was designed to enroll up to a maximum of 500 patients, with a prespecified interim analysis of data to assess efficacy upon enrollment of the first 200 patients.

According to Stryker, the independent Data Safety Monitoring Board recommended stopping study enrollment based on the preplanned interim review of data from the first 200 patients, which concluded that multiple prespecified stopping criteria were met. A final analysis of the data will be conducted upon completion of the remaining patient follow-up.

DAWN is an international, multicenter, blinded endpoint assessment, randomized study. The purpose of the study is to evaluate if mechanical thrombectomy with the Trevo Retriever plus medical management leads to superior clinical outcomes at 90 days as compared with medical management alone in appropriately selected patients treated 6 to 24 hours after last seen well. The Trevo Retriever indication within the DAWN trial is currently approved for investigational use only by the US Food and Drug Administration in the United States under an investigational device exemption study approval, advised Stryker.

Stryker's DAWN Trial Enrollment Ended After Interim Review Indicates Success

SHARE | E-MAIL | PRINT | BOOKMARK |

March 9, 2017—Stryker announced an early end to patient enrollment in the [DAWN](#) clinical trial comparing mechanical thrombectomy with the company's Trevo Retriever plus medical therapy versus medical therapy alone when initiated within 6 to 24 hours after time last known well. The study was designed to enroll up to a maximum of 500 patients, with a prespecified interim analysis of data to assess efficacy upon enrollment of the first 200 patients.

According to Stryker, the independent Data Safety Monitoring Board recommended stopping study enrollment based on the preplanned interim review of data from the first 200 patients, which concluded that multiple prespecified stopping criteria were met. A final analysis of the data will be conducted upon completion of the remaining patient follow-up.

DAWN is an international, multicenter, blinded endpoint assessment, randomized study. The purpose of the study is to evaluate if mechanical thrombectomy with the Trevo Retriever plus days as compared to medical therapy alone leads to superior clinical outcomes at 90 days as compared to medical therapy alone when initiated within 6 to 24 hours after time last known well. The trial is currently enrolling patients. Administrative and ethical review by the U.S. Food and Drug Administration approval, advised by the National Institute of Neurological Disorders and Stroke, is pending.

Primary Outcome Measures:

- Weighted modified Rankin Scale (mRS) score [Time Frame: 90 days]
- Stroke-related mortality [Time Frame: 90 days]

Trial record 1 of 11 for: DAWN stroke | Stroke

[Previous Study](#) | [Return to List](#) | [Next Study ▾](#)

Clinical Mismatch in the Triage of Wake Up and Late Presenting Strokes Undergoing Neurointervention With Trevo (DAWN)

This study is ongoing, but not recruiting participants.

Sponsor: Stryker Neurovascular

Information provided by (Responsible Party): Stryker Neurovascular

ClinicalTrials.gov Identifier: NCT02142283

First received: May 15, 2014
Last updated: April 6, 2017
Last verified: April 2017
[History of Changes](#)

[Full Text View](#) [Tabular View](#) [No Study Results Posted](#) [Disclaimer](#) [How to Read a Study Record](#)

Purpose

The purpose of the study is to evaluate the hypothesis that Trevo thrombectomy plus medical management leads to superior clinical outcomes at 90 days as compared to medical management alone in appropriately selected subjects experiencing an acute ischemic **stroke** when treatment is initiated within 6-24 hours after last seen well.

Condition	Intervention
Ischemic Stroke	Device: Trevo Thrombectomy Procedure Other: Medical Management

Study Type: Interventional
Study Design: Allocation: Randomized
 Intervention Model: Parallel Assignment
 Masking: Outcomes Assessor
 Primary Purpose: Treatment

Official Title: Diffusion Weighted Imaging (DWI) or Computerized Tomography Perfusion (CTP) Assessment With Clinical Mismatch in the Triage of Wake Up and Late Presenting Strokes Undergoing Neurointervention (DAWN)

Further study details as provided by Stryker Neurovascular:

Primary Outcome Measures:

- Weighted modified Rankin Scale (mRS) score [Time Frame: 90 days]
- Stroke-related mortality [Time Frame: 90 days]

Original Contribution

Adopting a Patient-Centered Approach to Primary Outcome Analysis of Acute Stroke Trials Using a Utility-Weighted Modified Rankin Scale

Napasri Chaisinanunkul, MD; Opeolu Adeoye, MD, MS; Roger J. Lewis, MD, PhD;
James C. Grotta, MD; Joseph Broderick, MD; Tudor G. Jovin, MD; Raul G. Nogueira, MD;
Jordan J. Elm, PhD; Todd Graves, PhD; Scott Berry, PhD; Kennedy R. Lees, MD;
Andrew D. Barreto, MD, MS; Jeffrey L. Saver, MD;
for the DAWN Trial and MOST Trial Steering Committees*

Background and Purpose—Although the modified Rankin Scale (mRS) is the most commonly used primary end point in acute stroke trials, its power is limited when analyzed in dichotomized fashion and its indication of effect size challenging to interpret when analyzed ordinally. Weighting the 7 Rankin levels by utilities may improve scale interpretability while preserving statistical power.

Methods—A utility-weighted mRS (UW-mRS) was derived by averaging values from time-tradeoff (patient centered) and person-tradeoff (clinician centered) studies. The UW-mRS, standard ordinal mRS, and dichotomized mRS were applied to 11 trials or meta-analyses of acute stroke treatments, including lytic, endovascular reperfusion, blood pressure moderation, and hemicraniectomy interventions.

Results—Utility values were 1.0 for mRS level 0; 0.91 for mRS level 1; 0.76 for mRS level 2; 0.65 for mRS level 3; 0.33 for mRS level 4; 0 for mRS level 5; and 0 for mRS level 6. For trials with unidirectional treatment effects, the UW-mRS paralleled the ordinal mRS and outperformed dichotomous mRS analyses. Both the UW-mRS and the ordinal mRS were statistically significant in 6 of 8 unidirectional effect trials, whereas dichotomous analyses were statistically significant in 2 to 4 of 8. In bidirectional effect trials, both the UW-mRS and ordinal tests captured the divergent treatment effects by showing neutral results, whereas some dichotomized analyses showed positive results. Mean utility differences in trials with statistically significant positive results ranged from 0.026 to 0.249.

Conclusions—A UW-mRS performs similar to the standard ordinal mRS in detecting treatment effects in actual stroke trials and ensures the quantitative outcome is a valid reflection of patient-centered benefits. (*Stroke*. 2015;46:00-00. DOI: 10.1161/STROKEAHA.114.008547.)

Key Words: blood pressure ■ stroke ■ stroke management ■ thrombolysis

The modified Rankin Scale (mRS) is the most widely used measure of outcome after acute ischemic stroke in both research clinical trials and national and local quality improvement registries. However, there is much debate on how best statistically to analyze the mRS.^{1,2} Approaches include simple dichotomization, sliding dichotomy or responder analysis, and ordinal or shift analysis.² The power of the mRS to detect treatment effects is often reduced when the scale is analyzed in dichotomized fashion, discarding substantial outcome

information. In the simple dichotomous approach, the 7 possible mRS scores are collapsed into just 2 health states, and the optimal point for dichotomization depends on timing of the intervention and the anticipated distribution of severity of illness and prognosis of enrolled subjects.¹ As data to guide selection of the most informative dichotomization are often incomplete, suboptimal selection may occur, missing a true treatment effect. Moreover, because they discard the preponderance of outcome information, dichotomized analyses

Received December 27, 2014; final revision received June 2, 2015; accepted June 8, 2015.

From the Department of Neurology and Comprehensive Stroke Center, University of California, Los Angeles (N.C., J.L.S.); Phyaithai Stroke Center, Department of Neurology, Phyaithai 1 Hospital, Bangkok, Thailand (N.C.); Departments of Emergency Medicine and Neurosurgery, Neuroscience Institute (O.A.) and Department of Neurology and Rehabilitation Medicine (J.B.), University of Cincinnati, OH; Department of Emergency Medicine at Harbor-UCLA Medical Center, Berry Consultants, LLC, Austin, TX (R.J.L.); Clinical Innovation Research Institute, Memorial Hermann Hospital-Texas Medical Center, Houston (J.C.G.); Department of Neurology, University of Pittsburgh Medical Center, PA (T.G.J.); Department of Neurology, Marcus Stroke and Neuroscience Center, Grady Memorial Hospital, Emory University, Atlanta, GA (R.G.N.); Department of Public Health Sciences, Medical University of South Carolina, Charleston (J.J.E.); Berry Consultants, LLC, Austin, TX (T.G., S.B.); Department of Stroke Research, University of Glasgow, Glasgow, United Kingdom (K.R.L.); and Department of Neurology, Stroke Division, University of Texas Health Science Center at Houston (A.D.B.).

*A list of all DWI/PWI and CTP Assessment in the Triage of Wake-Up and Late Presenting Strokes Undergoing Neurointervention (DAWN) Trial and Mode Selection Trial in Sinus Node Dysfunction (MOST) Trial Steering Committee contributors is given in the Appendix.

Guest Editor for this article was Eric E. Smith, MD.

Correspondence to Jeffrey L. Saver, MD, UCLA Comprehensive Stroke Center, 710 Westwood Plaza, Los Angeles, CA 90095. E-mail jsaver@mednet.ucla.edu
© 2015 American Heart Association, Inc.

Stroke is available at <http://stroke.ahajournals.org>

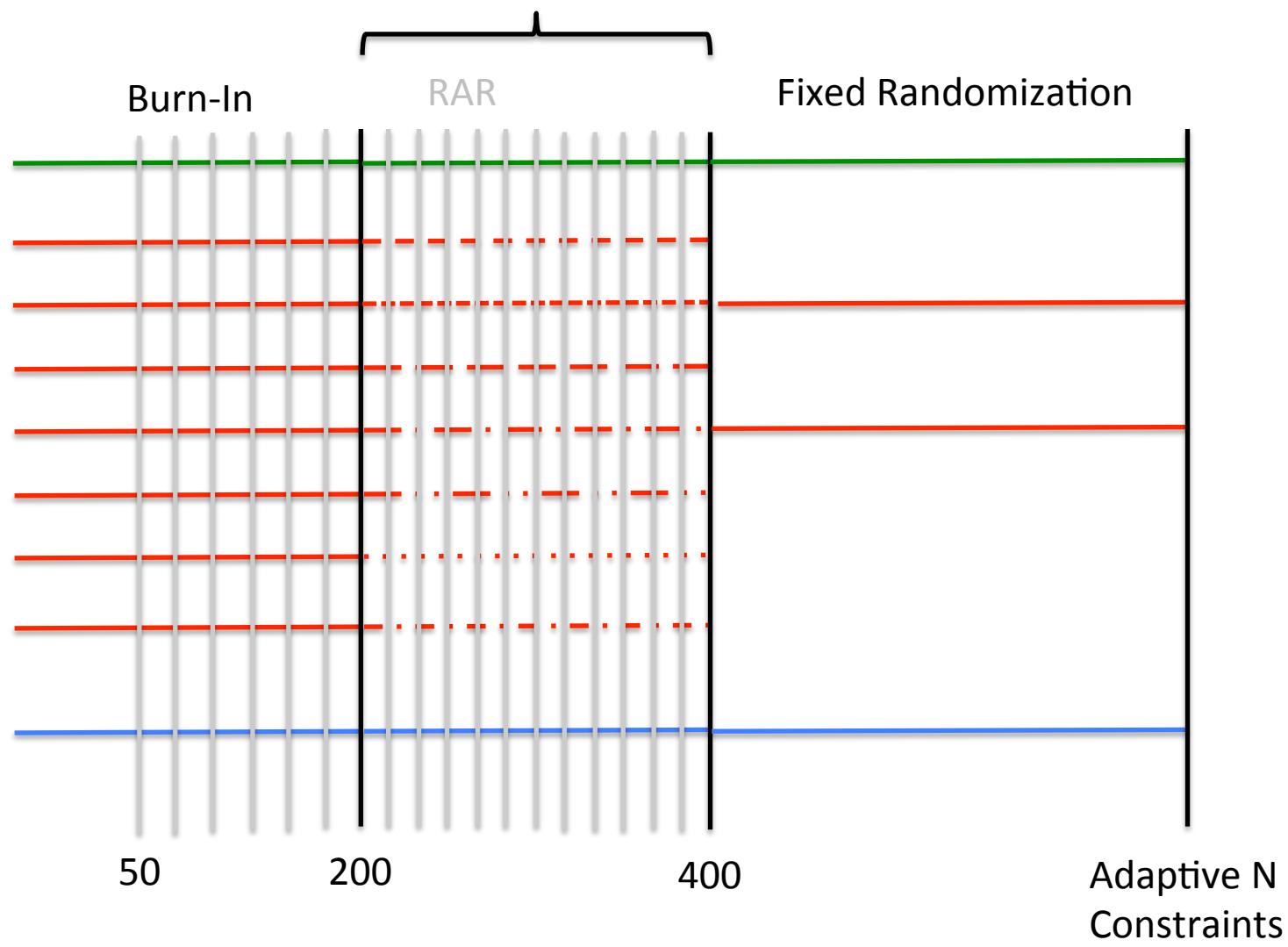
DOI: 10.1161/STROKEAHA.114.008547

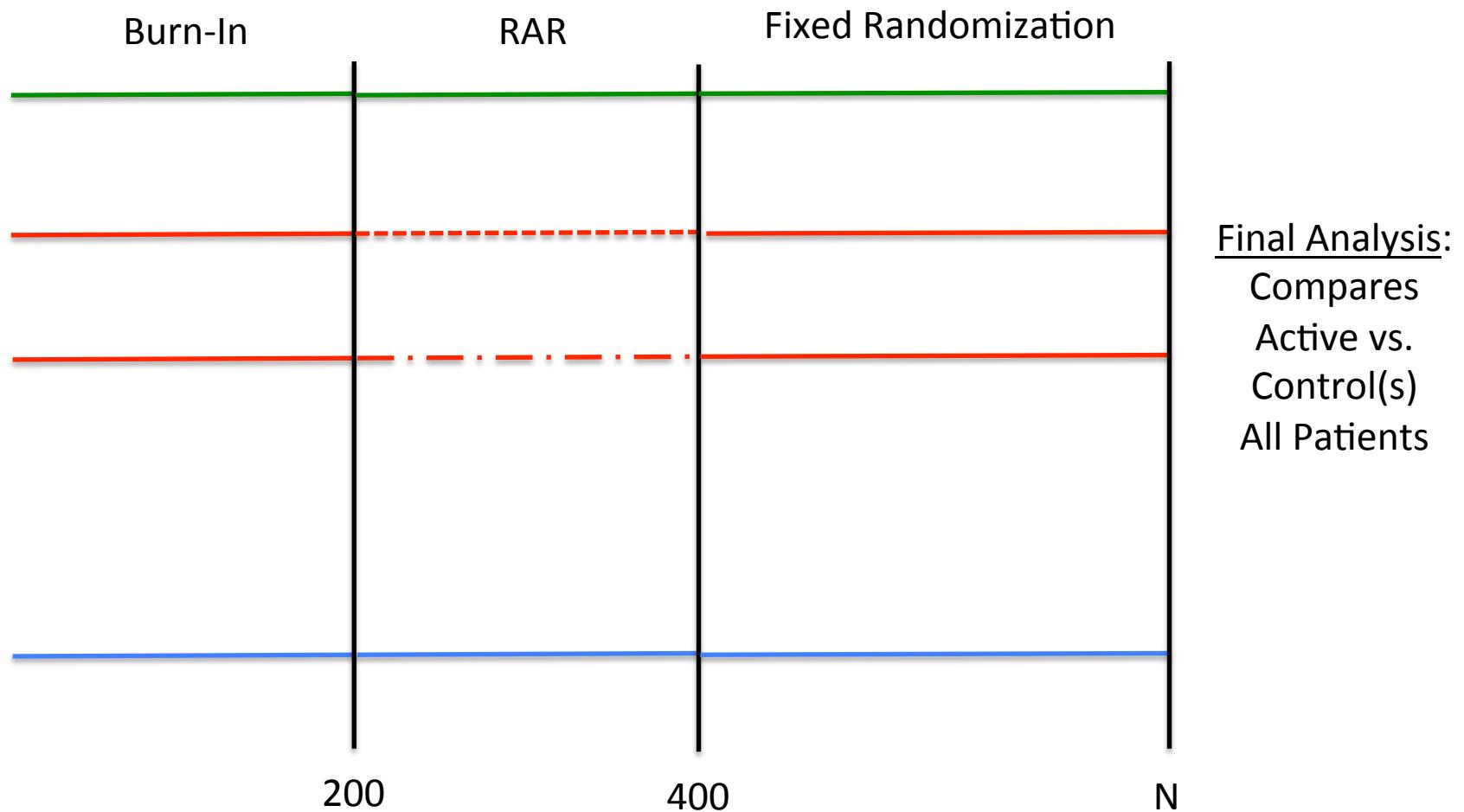
Downloaded from <http://stroke.ahajournals.org/> by guest on July 4, 2015

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

- Futility
- Go Part 2
- Forced @ 400





Bayesian Modeling

- Bayesian hierarchical-repeated-measures, dose-response, models for four endpoints
- Single utility function for value to sponsor connecting 4 endpoints
- Predictive probability of statistical success

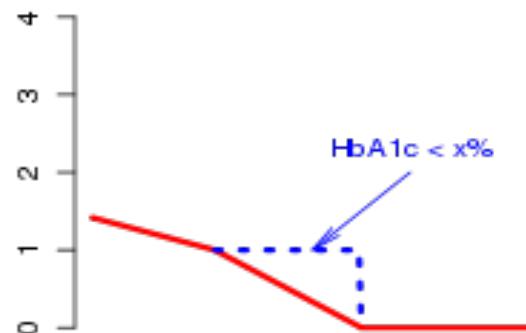
Utility of Drug/Dose

Utility for HbA1c



Dulaglutide minus Sitagliptin (%)
at 12 months

Utility for Weight



Dulaglutide minus Placebo (kg)
at 6 months

Utility for Pulse Rate



Dulaglutide minus Placebo (bpm)
at 6 months

Utility for Diastolic Blood Pressure



Dulaglutide minus Placebo (mmHg)
at 6 months

Go Decision

- Predictive probability of winning the phase III success endpoint for the *target dose* & probability of at least a 0.60 utility achieves threshold
- *Target dose* is most likely MaxU dose
- Bring two doses if predictive probability for a dose at least “2 levels” from target dose achieves a threshold

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
 - Several dozen dose scenarios

Diabetes II/III seamless

- 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

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

SYMPORIUM

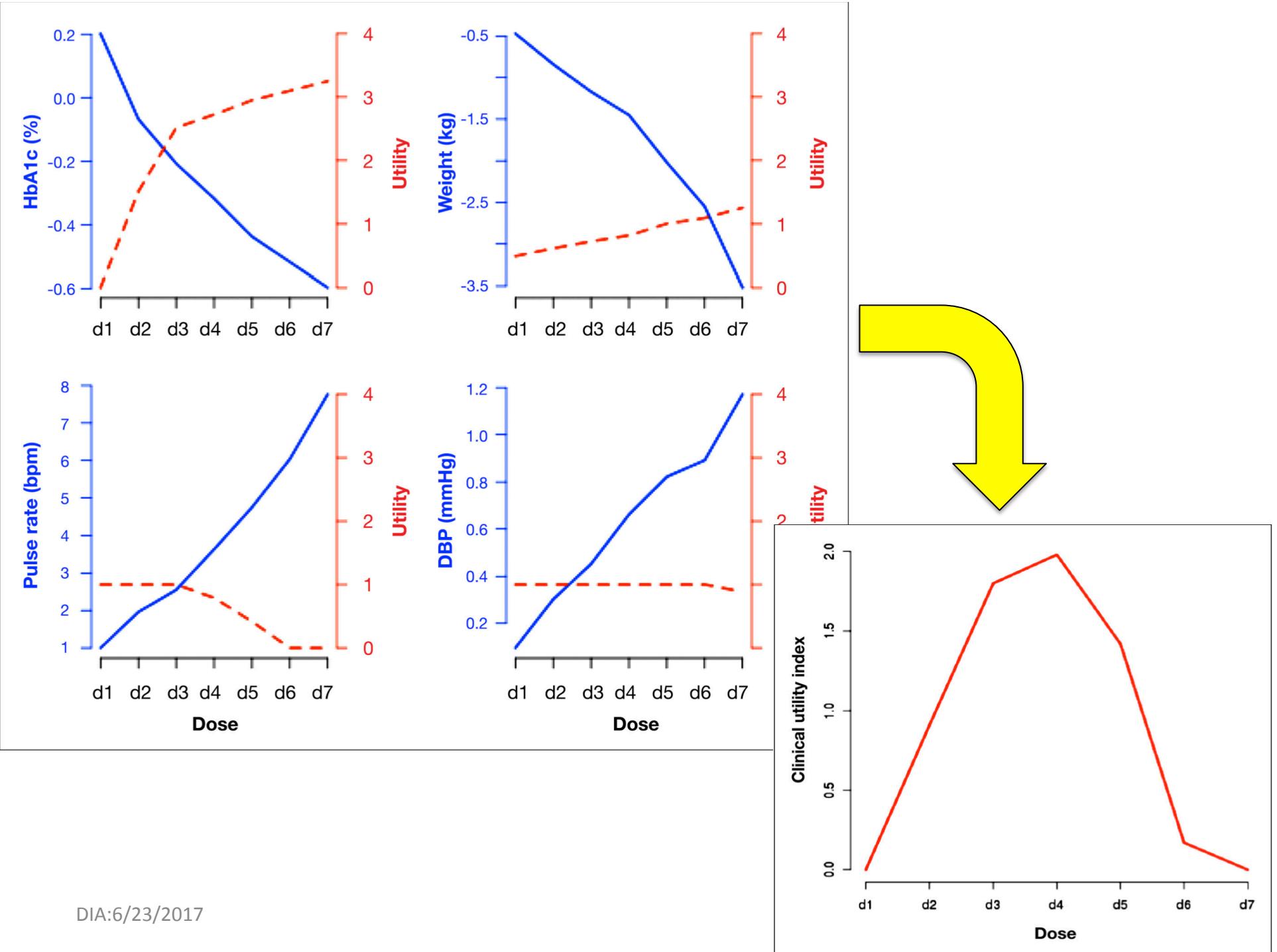
An Adaptive, Dose-Finding, Seamless Phase 2/3 Study of a Long-Acting Glucagon-Like Peptide-1 Analog (Dulaglutide): Trial Design and Baseline Characteristics

Mary Jane Geiger, M.D., Ph.D.¹ Zachary Skrivanek, Ph.D.¹ Brenda Gaydos, Ph.D.²
Jenny Chien, Ph.D.¹ Scott Berry, Ph.D.³ Donald Berry, Ph.D.^{3,4} and James H. Anderson, Jr., M.D.⁵

Abstract

Dulaglutide (dula, LY2189265) is a once-weekly glucagon-like peptide-1 analog in development for the treatment of type 2 diabetes mellitus. The first adaptive, dose-finding, inferentially seamless phase 2/3 study was designed to support the development of this novel diabetes therapeutic. The study is divided into two stages based on two randomization schemes: a Bayesian adaptive scheme (stage 1) and a fixed scheme (stage 2). Stage 1 of the trial employs an adaptive, dose-finding design to lead to a dual dose-selection decision or early study termination due to futility. If dose selection occurs, the study proceeds to stage 2 to allow continued evaluation of the selected dula doses. At completion, the entire study will serve as a confirmatory phase 3 trial. The final study design is discussed, along with specifics pertaining to the actual execution of this study and selected baseline characteristics of the participants.

J Diabetes Sci Technol 2012;6(6):PAGE NUMBERS



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

September 18, 2014 | By John Carroll

SHARE

64



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.

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

September 18, 2014 | By John Carroll

SHARE

64



An embattled Eli Lilly ([\\$LLY](#)) won a major battle today, gaining the FDA's approval to market dulaglutide for

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

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.



Lilly Diabetes President

UPDATE 2-Lilly revenue beats as diabetes drug sales rise

Tuesday, 31 Jan 2017 | 9:01 AM ET



- * Q4 sales of Trulicity, Humalog drugs beat estimates
- * Sales of newer drugs nearly triple
- * Shares marginally up at \$75.05 premarket

(Adds details, analysts comments, shares)

Jan 31 (Reuters) - Eli Lilly and Co's quarterly revenue beat analysts' estimates, driven by higher demand for its diabetes drug Humalog as well as its newer products such as Trulicity and Basaglar.

The company, whose earnings growth resumed in 2015 after three years of tumbling sales caused by competition from generic drugs, has been aggressively developing new drugs.

Eli Lilly said newer drugs generated sales of \$706.7 million in the fourth quarter ended Dec. 31, up from \$252.5 million a year earlier.

"Newly launched products, including Trulicity, Cyramza, Jardiance and

Trulicity, an injectable diabetes treatment that competes with Novo Nordisk's blockbuster Victoza, reported sales of \$337 million, easily beating consensus estimates of \$278 million, according to Evercore ISI.

Design/Project

- Bayesian modeling of 4 endpoints, repeated measures, and dose-response
- Frequent looks, response-adaptive randomization every 2 weeks, shifting to phase III, futility...
- Utilizes predictive probability for success and futility
- Utility function for characterizing ‘value’ to sponsor; maximize utility for dose-selection
- Balancing type I error with Bayesian machinery

Discussion

- Bayesian: What & Why
 - Bayesian Modeling
 - $\pi(\theta|x)$ and $f(x|\theta)$ really does matter
 - Prediction/Forecasting
 - Utilities
- The clash of philosophies in design
- Example using all of above!