Evaluating Probability of Success in Oncology Clinical Trials

Martin King, Ph.D.
Global Pharmaceutical R&D
Abbott

Abbott
A Promise for Life

Outline

- What is “success”?
- Probability of success vs. Power
- How does phase 2 affect the probability of success in phase 3?
- P(success) for binary data
- P(success) for time-to-event data
- Examples

Defining success

- Some possible definitions of success
  - P-value <0.05 vs. placebo
  - P-value <0.05 vs. placebo **with efficacy ≥ competing drug**
  - P-value <0.05 vs. placebo with efficacy ≥ competing drug **and better safety, tolerability, and convenience**

Probability of Success vs. Power

Typical phase 3 trial?

(From protocol) Determination of Sample Size

Assuming a significance level of 0.05 and an effect size of 0.30, the planned sample size of 176 subjects per group provides 80% power to detect a difference between drug and placebo.

Voicemail from clinical team: “We need a trial with 90% power, but we can’t afford to increase the sample size.”

Typical phase 3 trial?

(From protocol) Determination of Sample Size

Assuming a significance level of 0.05 and an effect size of **0.30**, the planned sample size of 176 subjects per group provides **80%** power to detect a difference between drug and placebo.

(From protocol) Determination of Sample Size

Assuming a significance level of 0.05 and an effect size of **0.35**, the planned sample size of 176 subjects per group provides **90%** power to detect a difference between drug and placebo.

But what is truly the probability of a successful trial?

Power vs. P(Success)

- Power is a **conditional** value
  - Choose an effect size
  - Power is the probability of statistical significance **if that is the true effect size**

- The probability of success is an **unconditional** value
  - P(success) is the weighted average of the power **across the range of possible effect sizes**
    - Expected value of power

See O'Hagan A, et al (Pharm Stat 2005;4:187-201) or Chuang-Stein C (Pharm Stat 2006;5:305-9) for more detailed discussions of the probability of success

How do we calculate the probability of success?

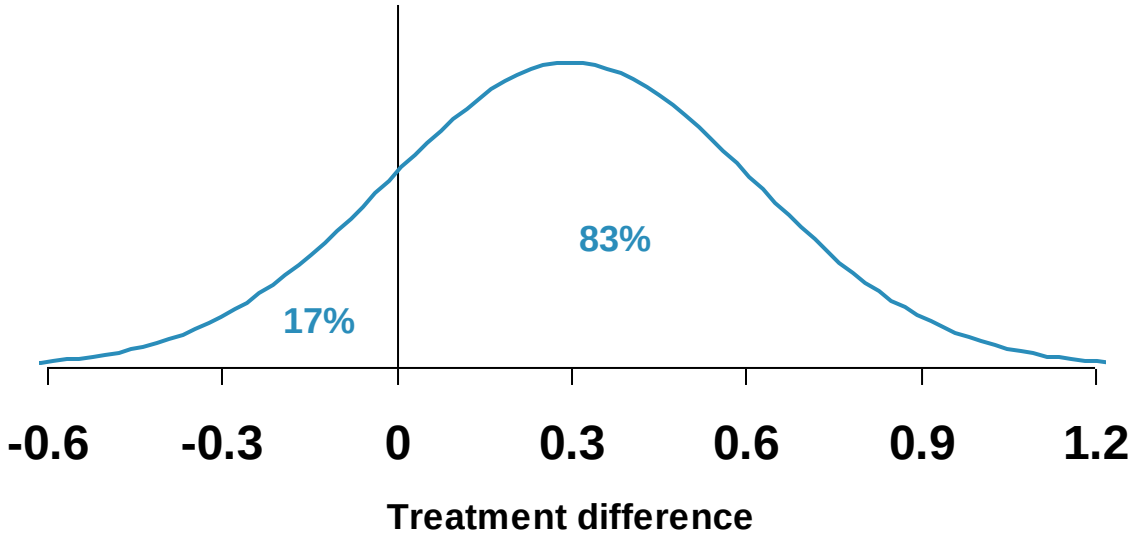
- Phase 2 trial – continuous endpoint
  - Drug vs. placebo, 20 subjects per arm
  - Mean difference is 0.3, SD is 1.0
  - Effect size = 0.3/1 = 0.3

- Naive approach to phase 3:
  - Effect size = 0.3
    - 176 subjects per group for 80% power
    - 235 subjects per group for 90% power

- But is 0.3 the right effect size?

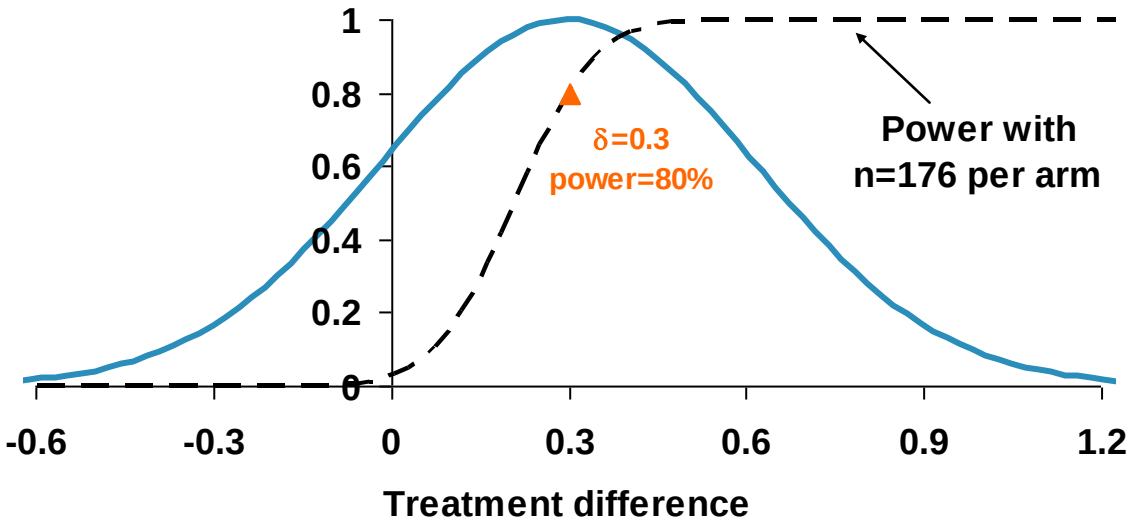
What do we know about the effect size?

- The phase 2 study implies a distribution of possible treatment differences
- (Of note, this is the posterior distribution of the true treatment difference, given the phase 2 study results and a uniform prior)



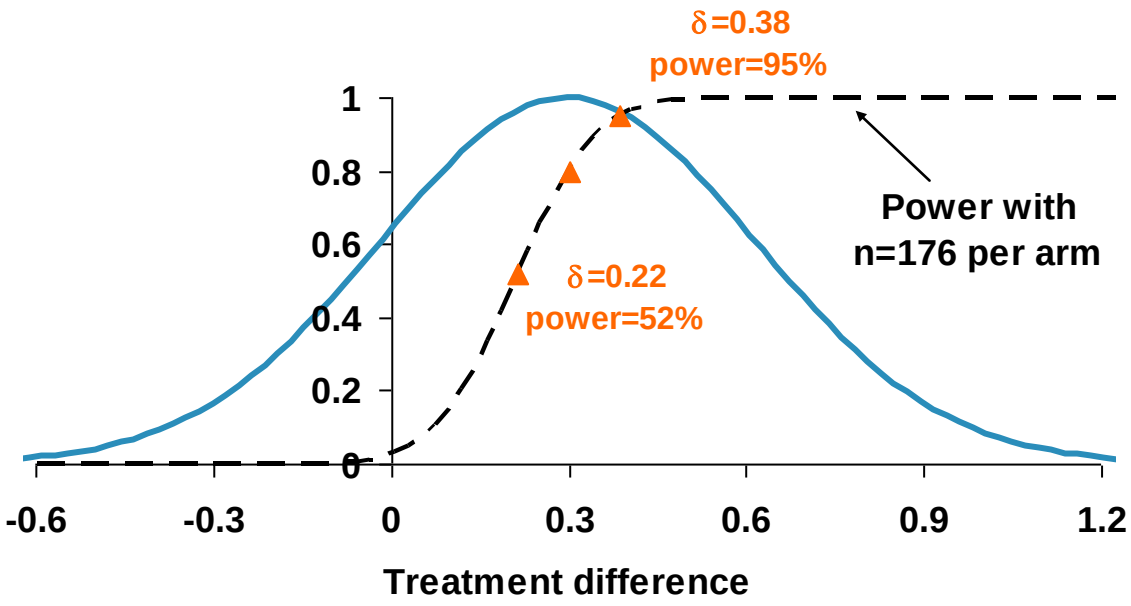
The central problem

- The power curve is asymmetric



The central problem

- The power curve is asymmetric

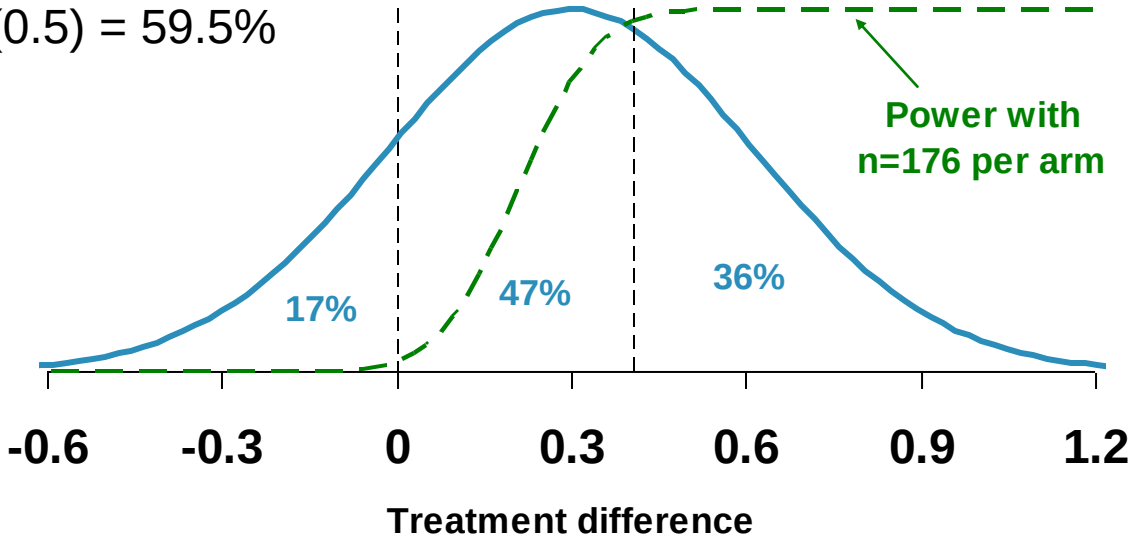


Calculating P(success) = expected power

- E(power) = $\int$ P(success|true diff) P(true diff|Ph2 diff) d(true diff)

- Crude numerical integration:
  - ~17% chance of ~0% power
  - ~36% chance of ~100% power
  - ~47% chance of ~50% power
  - 17%(0) + 36%(1) + 47%(0.5) = 59.5%

- Exact answer
  - 60.8%



Probability of success

d = observed difference in phase 2 study

s = observed standard deviation in phase 2 study

n_2 = number of subjects per group in phase 2 study

n_3 = planned number of subjects per group in phase 3 study

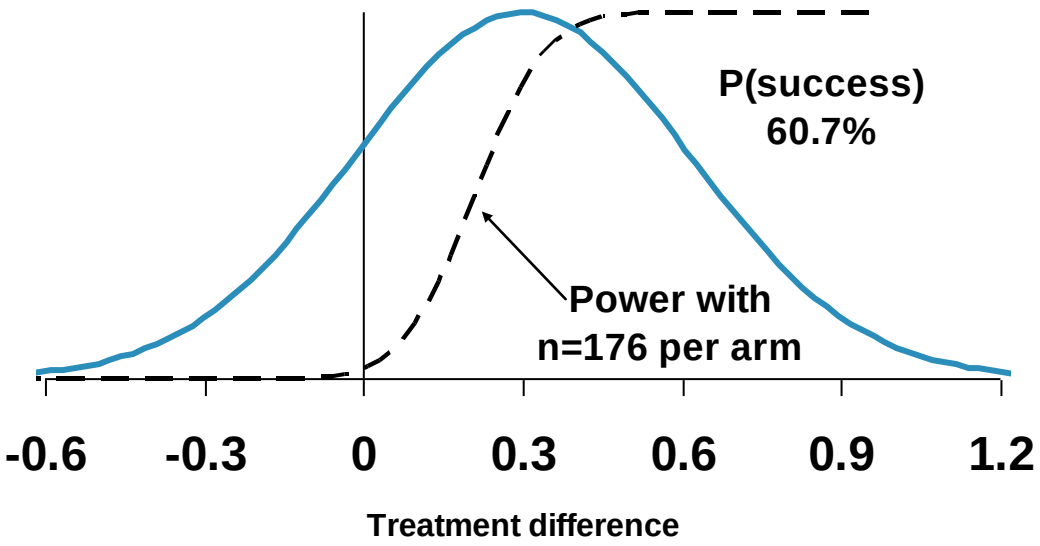
$$\text{Probability of success} = \Phi\left(\frac{d - 1.96s\sqrt{\frac{2}{n_3}}}{s\sqrt{\frac{2}{n_2} + \frac{2}{n_3}}}\right)$$

(1-sided significance test at $\alpha = 0.025$)

How does Phase 2 impact Probability of Success in Phase 3?

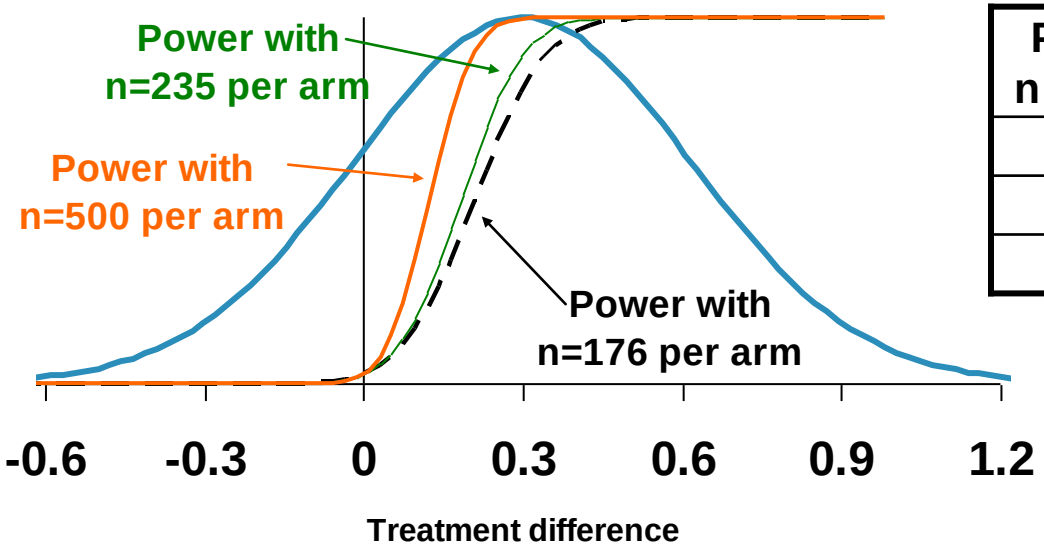
Improving the probability of success

- So we should add more subjects, right?



Improving the probability of success

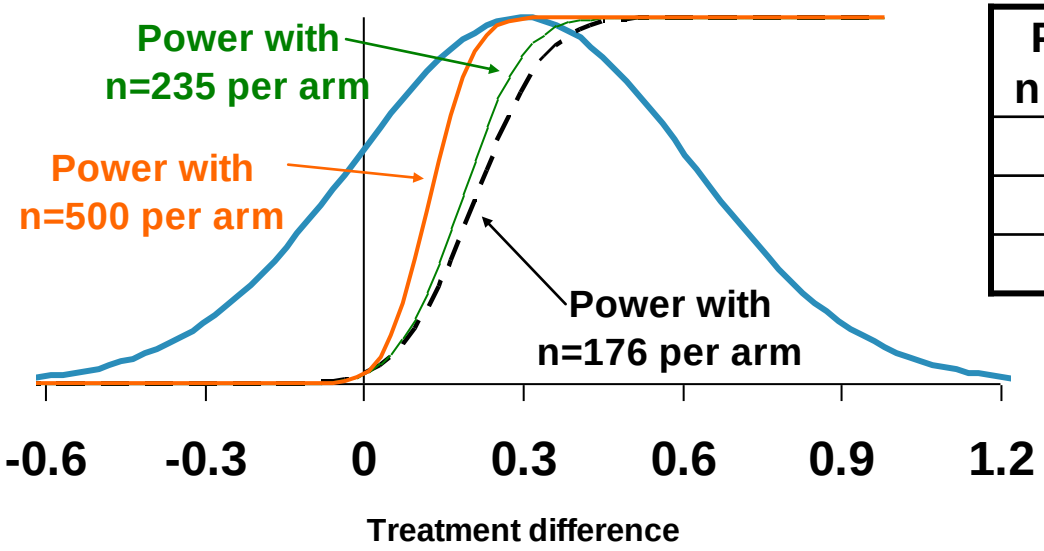
- So we should add more subjects, right?



Phase 3 n per arm	Power if true difference is 0.3
176	80%
235	90%
500	99.7%

Improving the probability of success

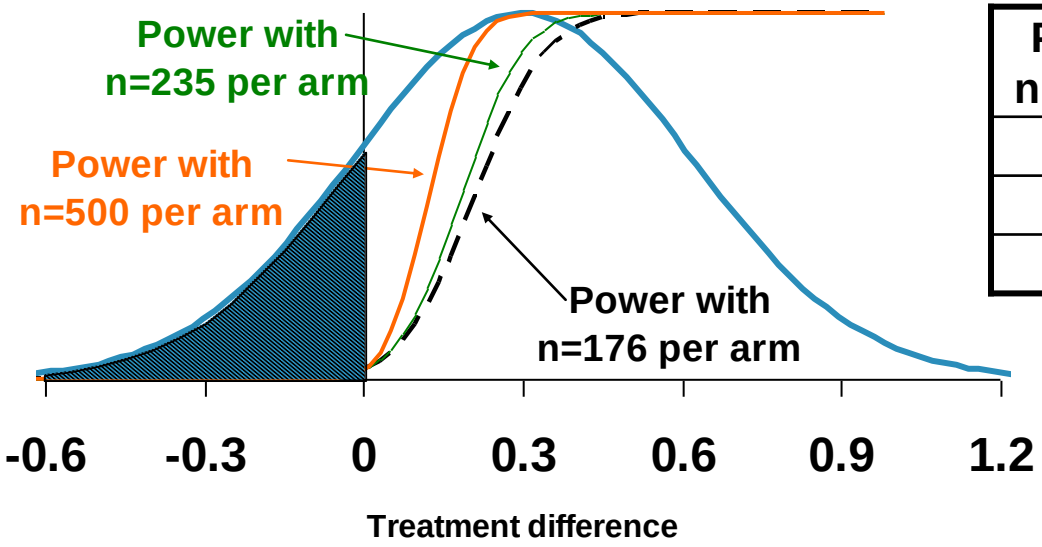
- So we should add more subjects, right?



Phase 3 n per arm	Power if true difference is 0.3	P(success)
176	80%	60.8%
235	90%	64.1%
500	99.7%	70.7%

Improving the probability of success

- So we should add more subjects, right?



Phase 3 n per arm	Power if true difference is 0.3	P(success)
176	80%	60.8%
235	90%	64.1%
500	99.7%	70.7%

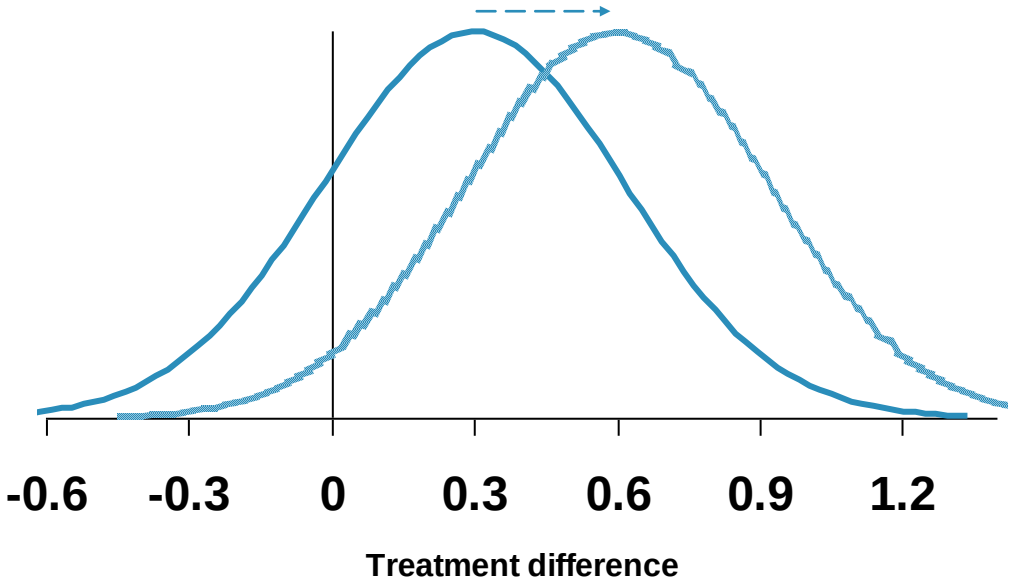
- The problem is not the power curves!
  - Too much blue curve at small or negative values

Improving the probability of success

- How do we move the blue curve?

  Get a better drug: effect size of 0.6 instead of 0.3.

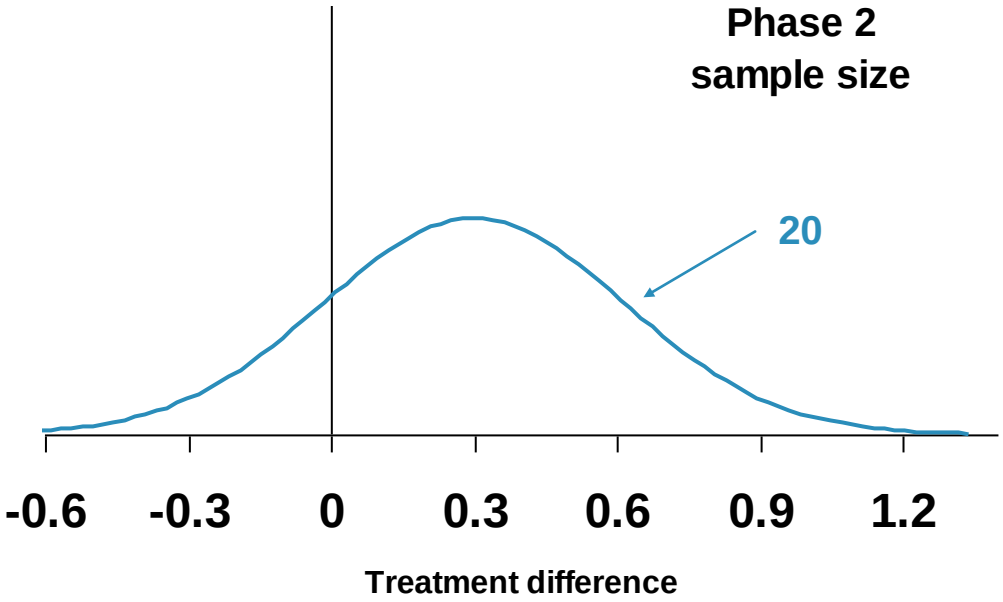
  With only n=88/arm in phase 3, P(success) is 81%



- More feasible: get a tighter estimate from Phase 2

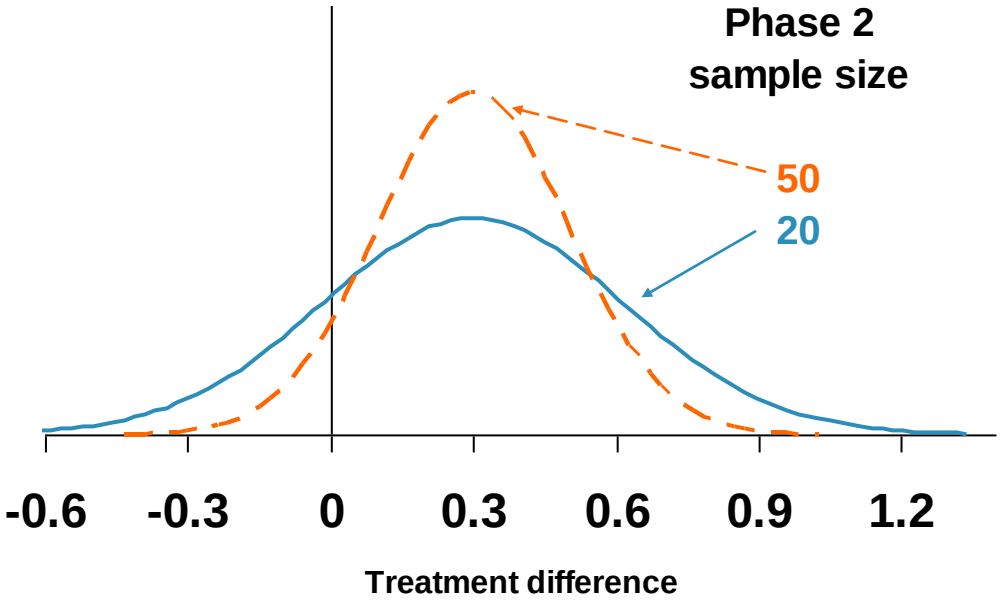
Improving the probability of success

- SD of treatment-difference curve is based on phase 2 sample size
  - ↑ phase 2 sample size → tighter estimate of effect size



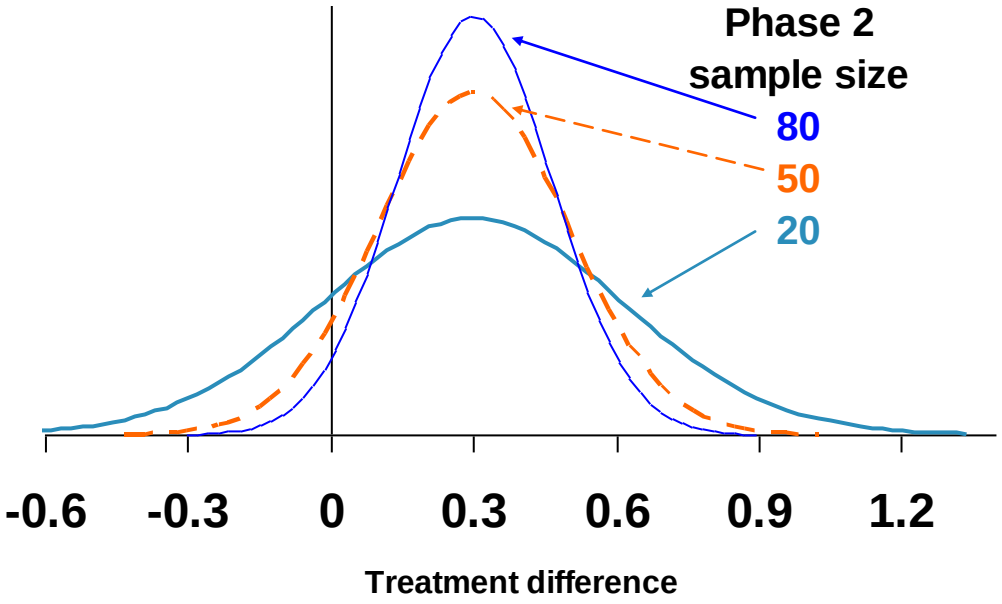
Improving the probability of success

- SD of treatment-difference curve is based on phase 2 sample size
  - ↑ phase 2 sample size → tighter estimate of effect size



Improving the probability of success

- SD of treatment-difference curve is based on phase 2 sample size
  - ↑ phase 2 sample size → tighter estimate of effect size



Phase 3: n=176 per arm

Phase 2 n per arm	Power if true difference is 0.3	P(success)
20	80%	60.8%
50	80%	65.6%
80	80%	68.4%

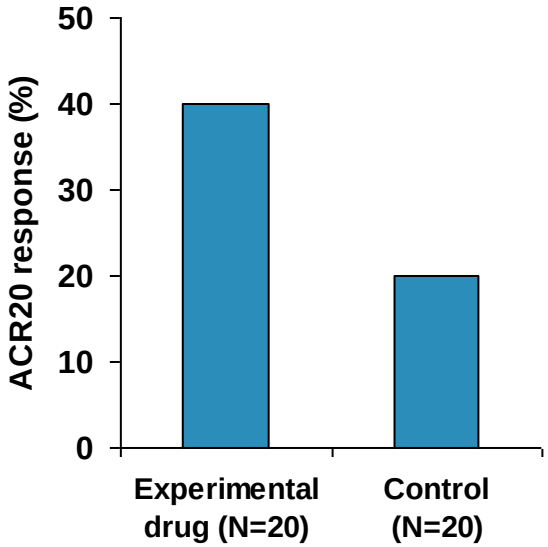
Phase 3: n=235 per arm

Phase 2 n per arm	Power if true difference is 0.3	P(success)
20	90%	64.1%
50	90%	70.6%
80	90%	74.3%

Probability of success for binary data

Binary data example: tumor response rate

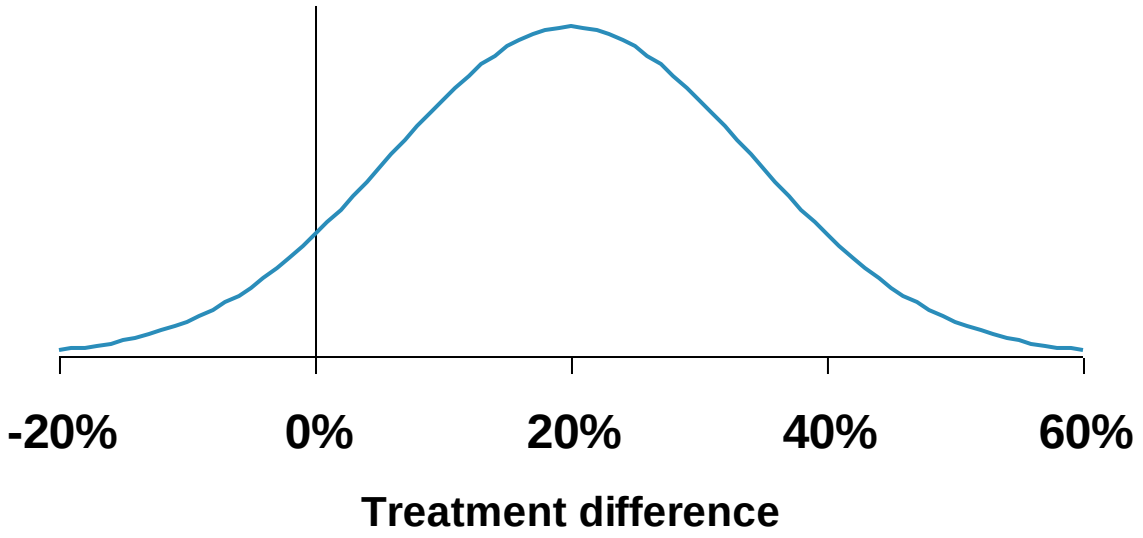
- Phase 2 study with 20 subjects per group
- Endpoint: Tumor response
- Results: Control 20%, Experimental drug 40%



- Naive phase 3 power calculation
  - Assume underlying response rates of 20% vs. 40%, 2-sided α=0.05
  - N=120/group provides 90% power
  - Does not account for uncertainty of response estimates

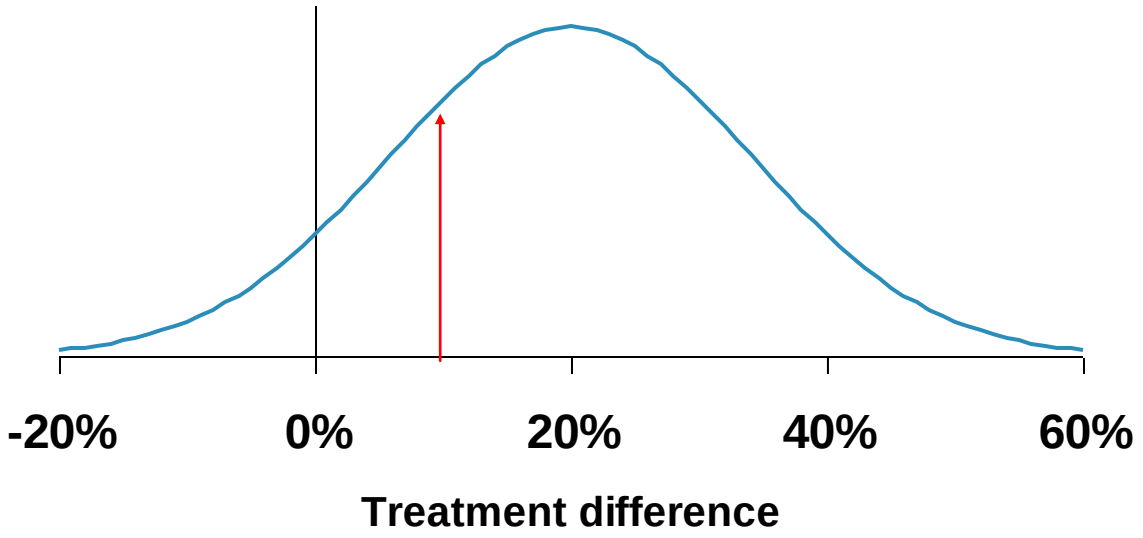
Complication with binary data

- Based on randomized trial, possible to construct posterior distribution for treatment difference
  - n=20 per arm, 20% vs. 40% response rate
- But problems arise computing expected value of power



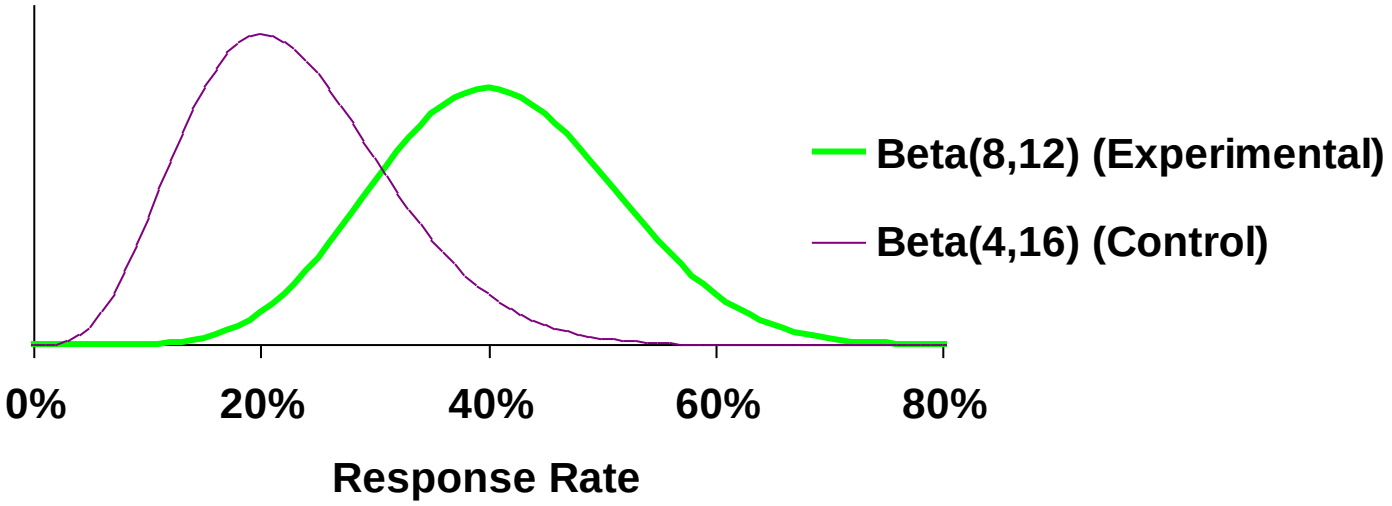
Complication with binary data

- To do the numerical integration, need to calculate power at each point across the distribution
  - Consider a specific point on the curve (difference of 10%, e.g.)
  - Since SD varies with specific rates, not possible to calculate power knowing only the difference in response rates
    - For a given sample size, power for 20% vs. 10% is higher than for 50% vs. 40%



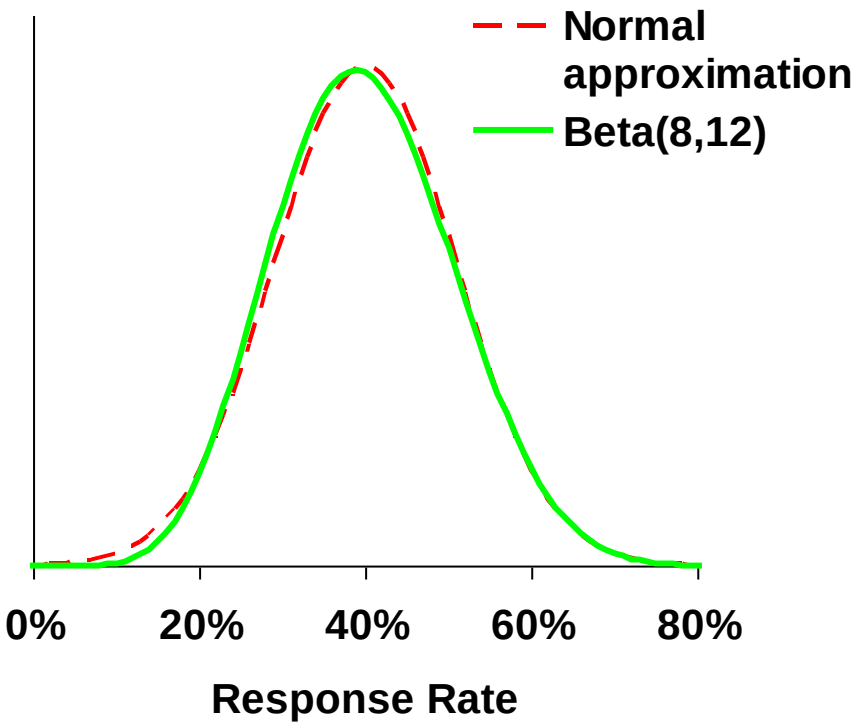
Uncertainty of response rate estimates

- How do we account for uncertainty of response rate estimates?
  - Consider Beta distribution to approximate the binomial for each group: Beta(α, β), where
    - α = # of responders
    - β = # non-responders
  - Control group (4 responders out of 20): Beta(4, 16)
  - Experimental group (8 responders out of 20): Beta(8, 12)

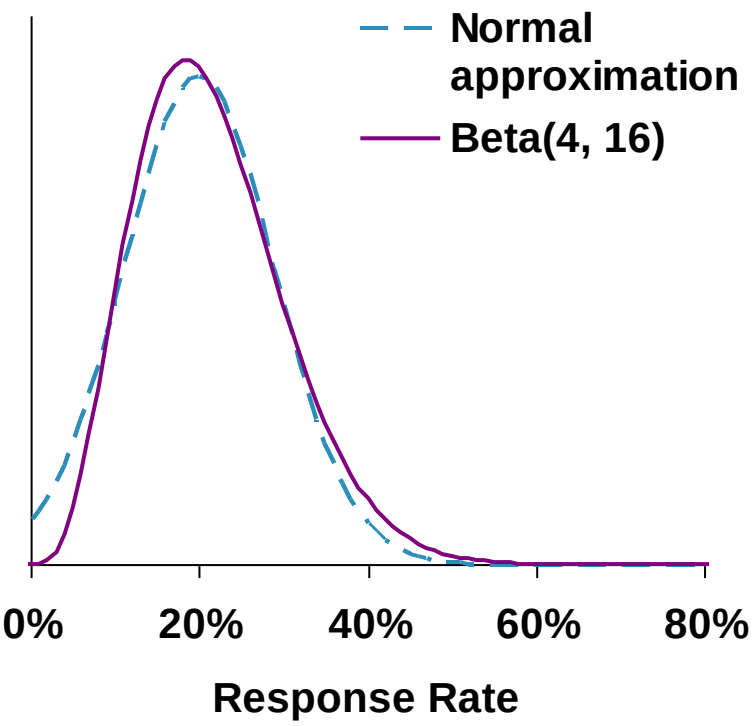


Beta distribution vs. normal approximation

Experimental group (8/20 responders)

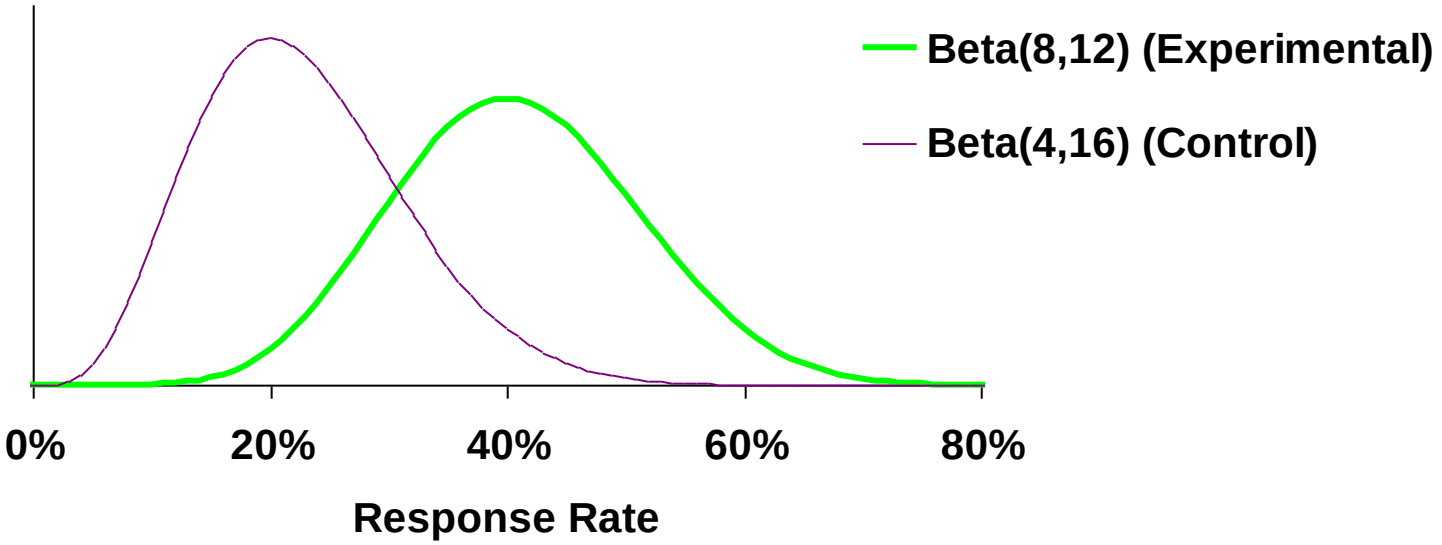


Control group (4/20 responders)



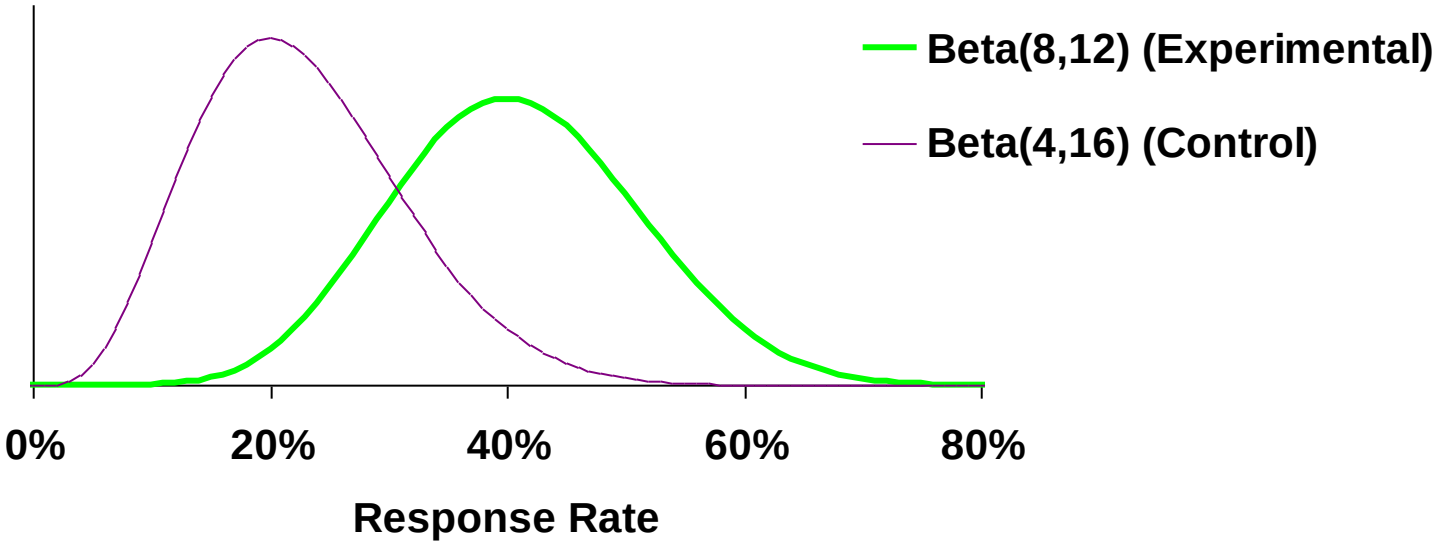
Simulations to compute P(success) for phase 3

- P(success) = P(p-value $\leq$0.05) = E(Power)
  1. Select response rate at random from each Beta distribution
  2. Calculate power based on selected response rates
  3. Repeat 1000 times (or 10,000, or 100,000)
  4. Compute average power across simulation runs



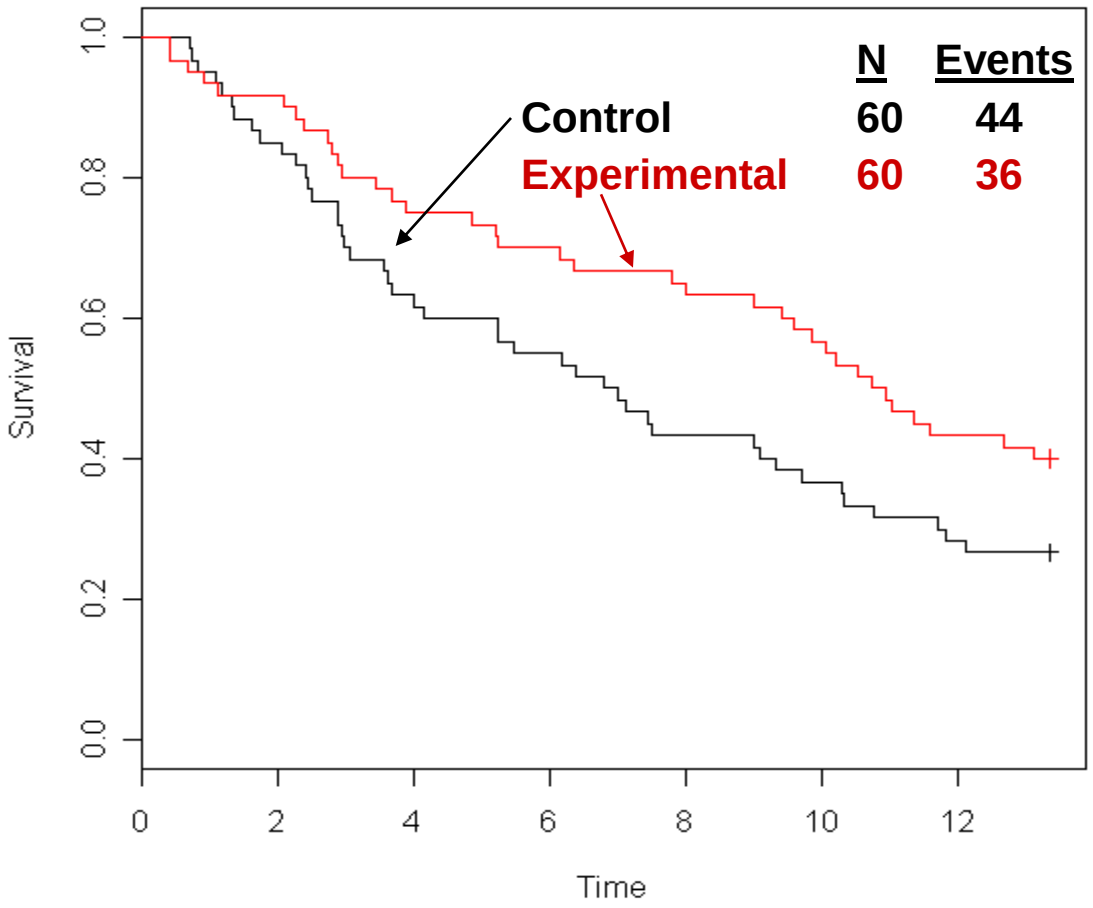
Simulation results

- Based on Phase 3 sample size of 120/group, P(success) = 66%
- Recall N=120/group provides 90% power in naive calculation that does not account for uncertainty in 20% vs. 40% response rates



Probability of success for time-to-event data

Phase 2 Study with time-to-event endpoint: Example



Median	Total follow-up
6.9	438
10.8	547

P-values

Log-rank	**0.06**
Cox PH	**0.061**

Hazard Ratio (95% CI)

0.656 (0.42, 1.02)

Probability of success for time-to-event data

- Simple implementation of probability of success – make use of the normal approximation for the log-hazard ratio

$$\log(hr) \sim N(\log(hr_2), 4/n_{e2})$$

- Where

hr_2 = observed hazard ratio in phase 2 study

n_{e2} = number of events in phase 2 study

Probability of success for time-to-event data

- Then the probability of success of the phase 3 trial (one-sided test at $\alpha = 0.025$) is

$$P(success) = P\left(\log(\hat{hr}) < -\sqrt{4/n_{e3}} \cdot 1.96\right) = \Phi\left(\frac{-\sqrt{4/n_{e3}} \cdot 1.96 - \log(hr_2)}{\sqrt{4/n_{e3} + 4/n_{e2}}}\right)$$

- Where

hr_2 = observed hazard ratio in phase 2 study

n_{e2} = number of events in phase 2 study

n_{e3} = planned number of events in phase 3 study

Example – Phase 3 time-to-event study

- Naive power estimate: If the true hazard ratio is 0.656, then 236 events provides 90% power
- Probability of success:

hr_2 = 0.656 = observed hazard ratio in phase 2 study
n_{e2} = 80 = number of events in phase 2 study
n_{e3} = 236 = planned number of events in phase 3 study

$$\text{P(success)} = \Phi\left(\frac{-\sqrt{4/236}\cdot 1.96 - \log(0.656)}{\sqrt{4/236 + 4/80}}\right) = 74.0\%$$

Probability of success for time-to-event data: a more general formulation

- Problem: may not always have a direct estimate of the hazard ratio
  - Single-arm phase 2 study
  - Historical data for phase 3 control arm
- Solution: Exponential – Inverse Gamma Model:
  - For exponential survival with parameter λ, let

$$\lambda \sim Inverse-gamma\,(a,b)$$

  - where a = number of events and b = total follow-up time.

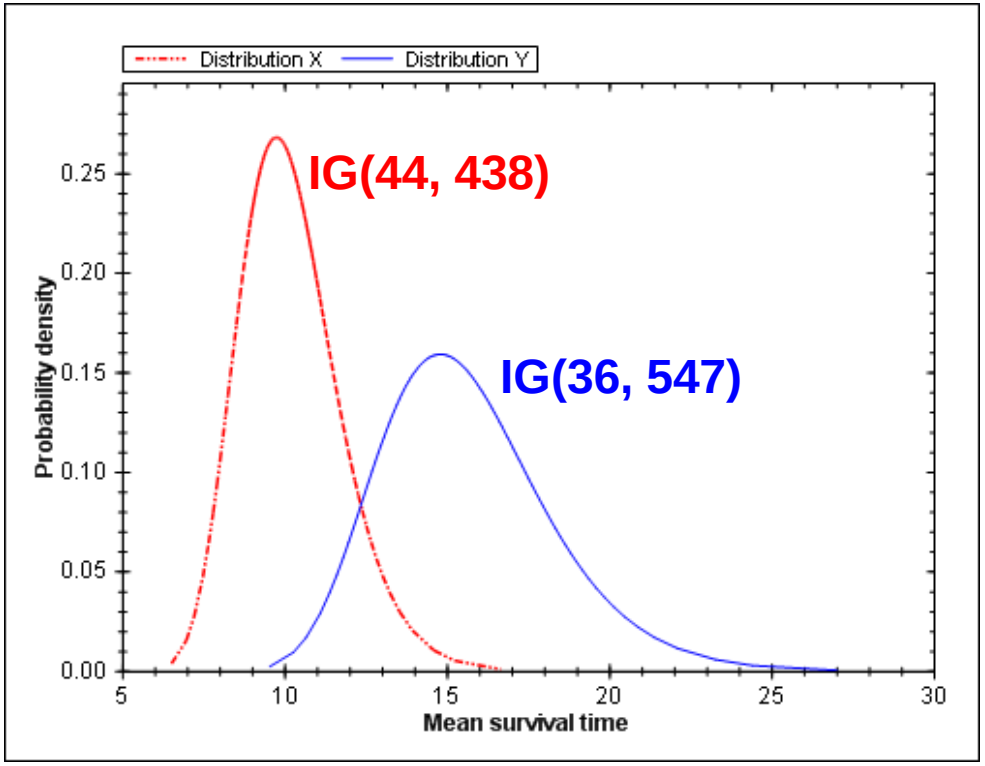
Inverse gamma

- Then for the two arms in the phase 2 study
  - $\lambda_{control} = IG(44,438)$

    Mean survival = 438/44 = 9.95
    Median = 9.95 * log(2) = 6.9

  - $\lambda_{experimental} = IG\ (36/547)$

    Mean survival = 547/36 = 15.2
    Median = 15.2 * log(2) = 10.5



Calculating P(success) from inverse-gamma model: Simulation Algorithm

1. Randomly draw a mean survival time from each inverse-gamma distribution.
2. “Enroll” patients into the study according to a certain accrual rate and randomize to experimental or control arm.
3. Draw event times randomly from the corresponding exponential distributions. Censor patients without events the end of the study.
4. Compare survival curves of experimental vs. control arms after the planned number of events is obtained.
5. Repeat steps 1 - 4 for a large number of replications K. Probability of success is calculated as number of times the trial results in a successful outcome / total number of replications K.

Summary of time-to-event data

- Time-to-event data have additional features and complexities compared to continuous (uncensored) data
- But the approach to assess the probability of success with time-to-event data is conceptually similar to that with other types of data
- The Bayesian framework used here can easily incorporate additional success criteria beyond the requirement of a p-value <0.05

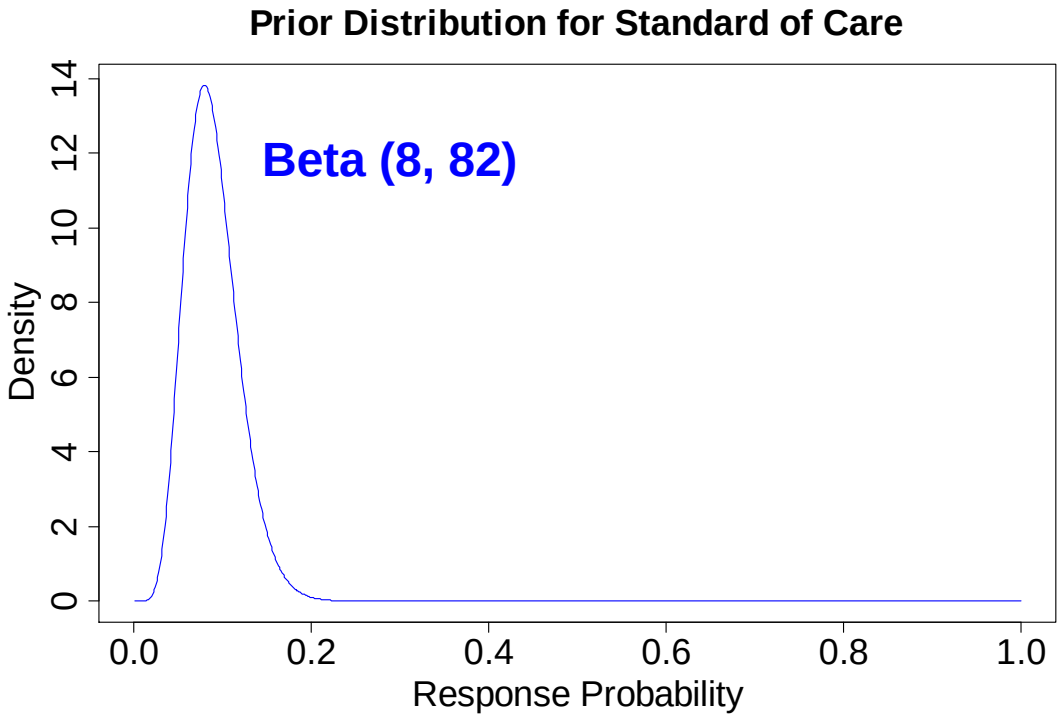
Examples

Example 1
Using P(success) to decide when to begin Phase 3

- 2nd-line or later treatment for a particular tumor type
- Uncontrolled Phase 2 study of experimental drug
- Endpoint: Response rate
- Standard of care: 9% response rate in prior uncontrolled trial of 90 subjects (8/90 subjects with partial response)
- Sample size: N=40
- Goal: determine whether to run a phase 3 study vs. the standard of care
  - Two phase 3 sample sizes considered: N=40/group or N=200/group

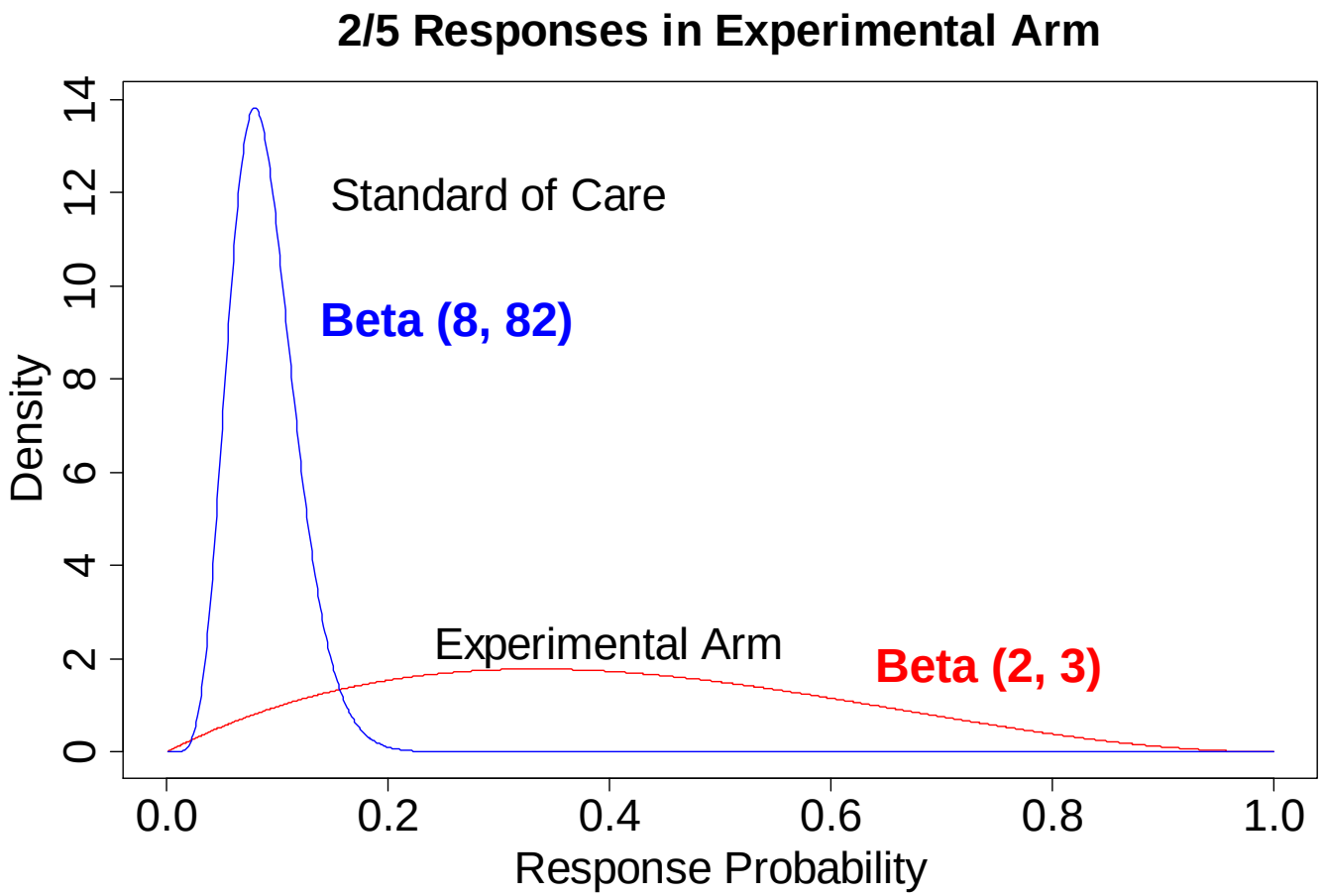
Standard of care treatment – historical data

- Standard of care: 9% response rate in prior uncontrolled trial of 90 subjects (8/90 subjects with partial response)
- Suggests a Beta(8, 82) distribution to characterize control arm
  - 8 responders
  - 82 non-responders



Establishing beliefs about response rate for experimental drug

- Suppose 2 responders in first 5 subjects



What do we know after 2/5 responses?

- Naive power calculation
  - N=40/arm in phase 3 study provides 86% power if the true rates are 9% (control) vs. 40% (experimental)
- To get probability of success ($p<0.05$ in phase 3 study), simulate:
  - Select response rates from Beta(8,82) and Beta(2,3) distributions
  - Compute power based on N=40/group
  - Repeat a large number of times calculate average power
- With N=40/arm, P(success) = 68%
- With N=200/arm, P(success) = 99%

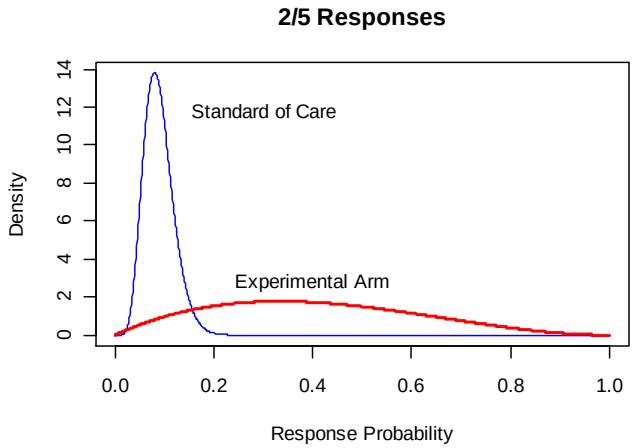
P(success) of phase 3 study after 5 subjects in phase 2 study

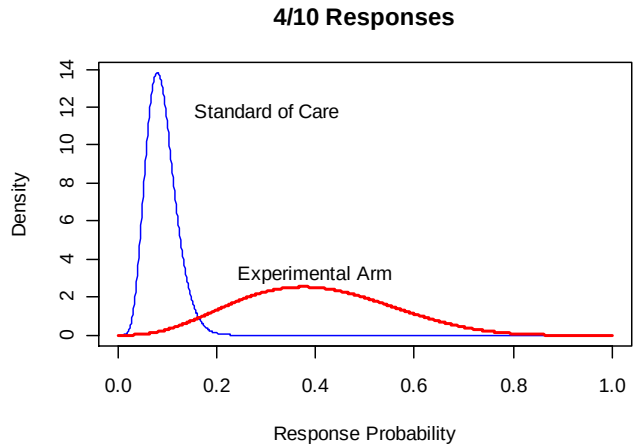
- P(success)
  - Select response rate at random from each Beta distribution
  - Calculate power based on selected response rates
  - Repeat 1000 times (or 10,000, or 100,000)
  - Compute average power across simulation runs

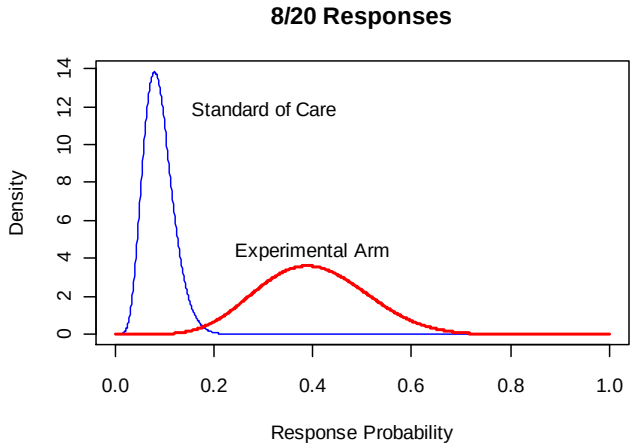
Phase 2 outcome (# of responses out of 5 subjects)	P(superiority) in phase 3 study at n=40/arm	P(superiority) in phase 3 study at n=200/arm
1 (20%)	0.28	0.49
2 (40%)	0.68	0.87
3 (60%)	0.91	0.99

Strength of evidence vs. sample size

- Our beliefs about the true response rate for the experimental drug get stronger with more subjects
  - 40% response rate based on 5, 10, and 20 subjects →







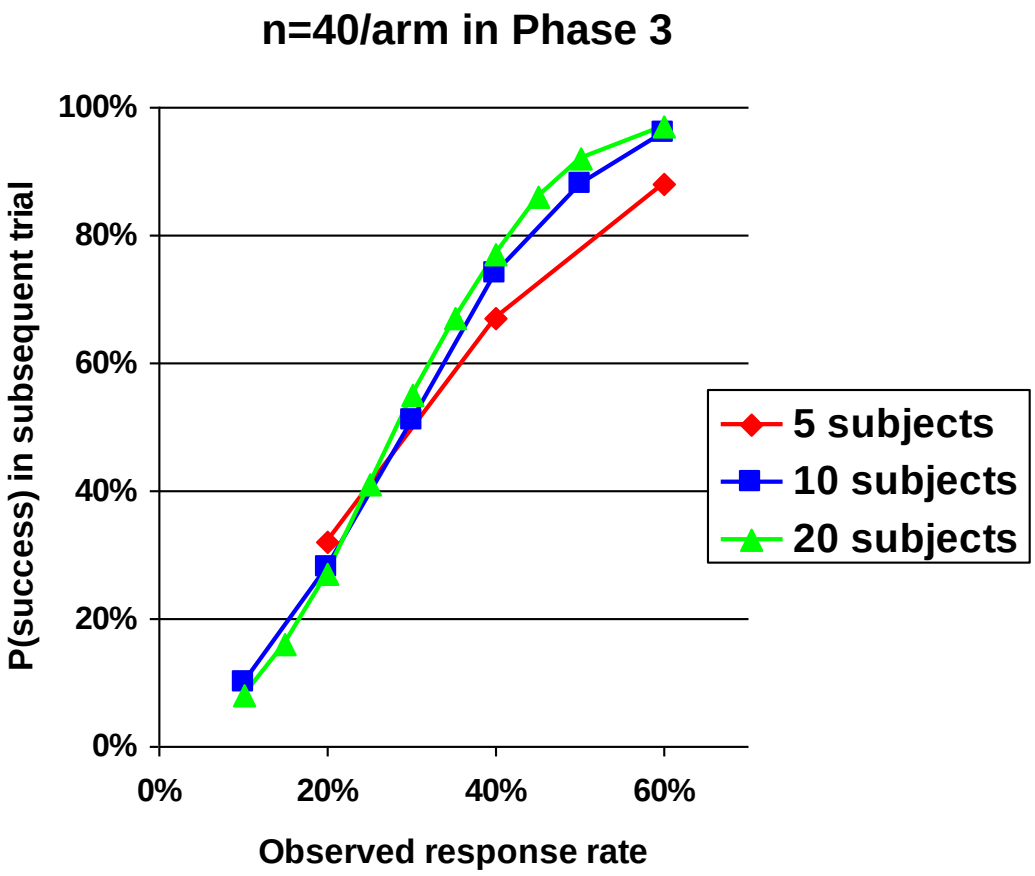
P(success) of phase 3 study after 10 subjects in phase 2 study

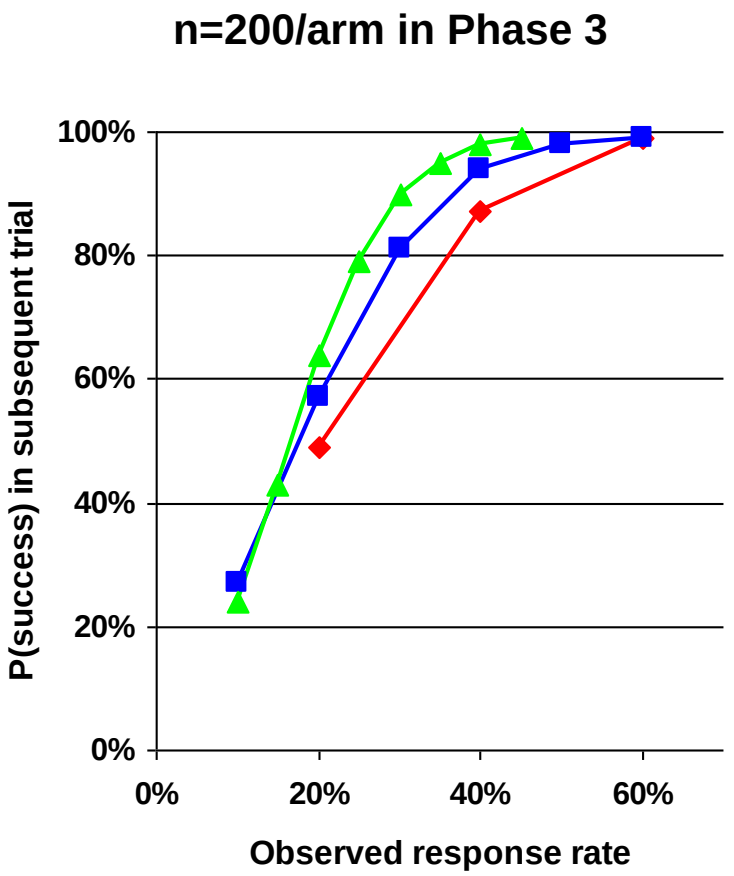
Phase 2 outcome (# of responses out of 10 subjects)	P(superiority) in phase 3 study at n=40/arm	P(superiority) in phase 3 study at n=200/arm
1 (10%)	0.10	0.27
2 (20%)	0.28	0.57
3 (30%)	0.51	0.81
4 (40%)	0.74	0.94
5 (50%)	0.88	0.98
6 (60%)	0.96	0.99

P(success) of phase 3 study after 20 subjects in phase 2 study

Phase 2 outcome (# of responses out of 20 subjects)	P(superiority) in phase 3 study at n=40/arm	P(superiority) in phase 3 study at n=200/arm
2 (10%)	0.08	0.24
3 (15%)	0.16	0.43
4 (20%)	0.27	0.64
5 (25%)	0.41	0.79
6 (30%)	0.55	0.90
7 (35%)	0.67	0.95
8 (40%)	0.77	0.98
9 (45%)	0.86	0.99

P(success) in Phase 3 by Phase 2 response rate





Example 2
Using P(success) to evaluate a development plan

- Randomized phase 2 study is about to start
  - Primary endpoint: overall survival, 30% improvement considered clinically meaningful
  - Number of events in phase 2 study: 40 vs. 60. vs. 80 vs. 100 ?
  - Company willing to run a 460-event phase 3 study (80% power for a true improvement of 30%) if P(success) is high enough
  - What is a “high enough” probability of success?

Possible outcomes

Outcome	Considerations
Win in phase 2	Probability, date of approval
Lose in phase 2 (stop development)	Probability, study cost, P(type II error) (stopping development if drug actually works)
Continue to phase 3 and win	Probability, date of approval
Continue to phase 3 and lose	Probability, study cost

P(success) for the phase 3 study based on phase 2 results

- Based on 460-event phase 3 trial:

Observed % improvement	Observed HR	P(success) in Phase 3 for given observed % improvement and given # of events in Phase 2					
		20	40	60	80	100	120
25.0%	0.8000	0.535	0.549	0.558	0.566	0.573	0.578
30.0%	0.7692	0.569	0.595	0.614	0.629	0.641	0.651
35.0%	0.7407	0.601	0.639	0.665	0.686	0.703	0.716
40.0%	0.7143	0.632	0.679	0.712	0.737	0.757	0.773
45.0%	0.6897	0.660	0.717	0.754	0.782	0.804	0.821
50.0%	0.6667	0.687	0.750	0.791	0.821	0.844	0.861

- What should the rule be to move into phase 3?
  - P(success) >80%?
  - P(success) >75%?
  - P(success) >60%?

Probability of each outcome in phase 2 or phase 3

- Assumptions
  - Conduct phase 3 study if P(success) is at least 75%
  - true HR is 0.7692 (30% improvement)
  - 40-event phase 2 study
  - 460-event phase 3 study

P(win in Phase 2)	= P(observed HR < 0.536) = 0.13
P(stop after Phase 2)	= P(observed HR ≥ 0.667) = 0.67
P(continue to Phase 3 and win)	= 80% * (1 – 0.13 – 0.67) = 0.16
P(continue to Phase 3 and lose)	= 20% * (1 – 0.13 – 0.67) = 0.04

Probability of each outcome by phase 3 decision rule and true underlying hazard ratio

- Phase 2 design: 40-event study

Outcome	Run phase 3 if P(success) is >75%			Run phase 3 if P(success) is >60%		
	True HR			True HR		
	1.000	**0.769**	**0.667**	**1.000**	**0.769**	**0.667**
Win in phase 2	0.025	0.13	0.25	0.025	0.13	0.25
Stop after phase 2	0.90	0.67	0.50	0.80	0.51	0.33
Win in phase 3	0.00	0.16	0.25	0.00	0.29	0.42
Lose in phase 3	0.07	0.04	0.00	0.17	0.07	0.00

Probability of each outcome by phase 3 decision rule and true underlying hazard ratio

- Phase 2 design: 100-event study

Outcome	Run phase 3 if P(success) is >75%			Run phase 3 if P(success) is >60%		
	True HR			True HR		
	1.000	**0.769**	**0.667**	**1.000**	**0.769**	**0.667**
Win in phase 2	0.025	0.26	0.53	0.025	0.26	0.53
Stop after phase 2	0.95	0.64	0.36	0.88	0.45	0.20
Win in phase 3	0.00	0.09	0.12	0.00	0.23	0.27
Lose in phase 3	0.02	0.02	0.00	0.09	0.06	0.00

Assigning value to each outcome

- Model inputs
  - 40-event phase 2 study
  - true HR = 0.7692 (30% improvement)
- Model output

- Study cost assumptions:
  - 40-event study = 4 MM
  - 100-event study = 10 MM
  - Phase 3 study = 50 MM

Outcome	P(outcome) (60% rule)	P(outcome) (75% rule)	Timing of outcome	Value
Win in phase 2	0.13	0.13	Approval in 2Q2014	XXX MM
Stop after phase 2	0.51	0.67	Study ends in 2Q2013	–4 MM
Win in phase 3	0.29	0.16	Approval 2Q2018	YYY MM
Lose in phase 3	0.07	0.04	Study ends 1Q2017	–54 MM

Expected value (60% rule) = 0.13(XXX) + 0.51(–4) + 0.29(YYY) + 0.07(–54)
Expected value (75% rule) = 0.13(XXX) + 0.67(–4) + 0.16(YYY) + 0.04(–54)

Extensions

- How do different beliefs about the drug's efficacy affect expected value?
  - Individual 1 believes there's 50% chance the drug has no efficacy (HR=1.0) and a 50% chance the drug gives a 30% improvement (HR=0.769)
  - Individual 2 believes there's 75% chance the drug has no efficacy (HR=1.0) and a 25% chance the drug gives a 30% improvement (HR=0.769)

- Calculate weighted average of the expected values for HR=1.00 and HR=0.769 and compare between individuals

Closing remarks

Complications

- Phase 3 is just like phase 2, except
  - Different year
  - Different sites
  - Different dose?
  - Different design
  - Different endpoint
  - Different formulation
  - Different inclusion criteria
  - Different statistical analysis
- Furthermore, development programs rarely consist of a single phase 2 study and a single phase 3 study

Conclusions

- Remember that
  - Power is a conditional value (more importantly, remind your clinical team)
  - The foundation for success in phase 3 is built in phase 2
  - The optimal probability of success may or may not be the familiar 80% or 90%

References

O'Hagan A, et al. Pharm Stat 2005;4:187-201.

Chuang-Stein C. Pharm Stat 2006;5:305-9.