Group Sequential Designs (and related topics)

Berry Adaptive Design and FACTS Webinar

Kert Viele

X/Twitter @KertViele, LinkedIn

Berry Consultants

Statistical Innovation

Outline

The Basics of Group Sequentials (GSDs)

- What problem are we solving?
- How do GSDs work? (success and futility)
- What advantages might occur in practice?

FACTS implementation

- Interim Schedules
- Futility
- Performance

Advanced Topics

- Are GSDs biased?
- What about delayed outcomes?
  - Goldilocks trials (including FACTS)
  - Use of longitudinal information

Regulatory

- Does FDA accept GSD? What kind?
- Importance of first interim timing
- “Information leakage” and operational bias
  - What can I tell from a press release?

What problem does a group sequential solve?

- At least two important settings for a group sequential

- Historically, GSDs recommended to save sample size?
  - A 90% powered trial is an insurance policy against bad luck
  - If we power for effect X, trial successful when we observe 0.6X
  - If we observe X or better, we obtain convincing evidence earlier.

- With random data, convincing evidence can occur at a random time. Why go longer than you need to?

Uncertainty in Treatment Effect for Power

- Suppose we had uncertainty about μ prior to the trial
  - Let's be honest here, we always have uncertainty….
- Consider just small uncertainty, μ=0.15 or μ=0.20
  - for μ=0.20, suppose need N=263
  - for μ=0.15, need N=467
  - those are VERY different.
- If we….
  - use N=263, ok for μ=0.20, but only 68% power for μ=0.15
  - use N=467, powered for μ=0.15, but bigger trial than needed for μ=0.20
- Good to have a trial which behaves well for both μ
  - Flexible sample sizes, appropriate for range of anticipated effects

Uncertainty in Treatment Effect for Power

- Suppose we had uncertainty about μ prior to the trial
  - Let's be honest here, we always have uncertainty….

- Consider just small uncertainty, μ=0.15 or μ=0.20
  - for μ=0.20, suppose need N=263
  - for μ=0.15, need N=467
  - those are VERY different.

Use N=263?
Good power for μ=0.20
But 68% power for μ=0.15

Use N=467?
Powered for both μ
Wasteful for μ=0.20

OR….
Flexible Sample Sizes
Look at both N

Basic Idea

- Perform interim analyses
  - At prespecified N (N_1, N_2, N_3, etc.) have a third party look at the data
  - If the data is "sufficiently good" (more later) declare efficacy, otherwise continue to the next interim analysis

- This allows
  - the trial may stop when the data indicate the question is answered
  - if μ is large, the trial is likely to stop with a smaller sample size
  - if μ is small, the trial can be big enough to detect it

- Key complexity
  - Looking at the data multiple times creates a multiplicity
  - We can't test $p<0.025$ multiple times, or the total probability of type 1 error will exceed 2.5%

A group sequential design

- K interim analyses at $N_1,\ldots,N_K$ (N_K is the maximal size)

- Reject H_0 whenever $p_k < \alpha_k$
  - p_k is the nominal p-value (usual calculation) at interim k
  - α_k are user selected, but must satisfy
    - Pr(any type 1 error) = 2.5% (or other needed overall alpha)

- Note the interim results are correlated
  - the first N_2 observations contain the first N_1 observations
  - The α_k values may sum to more than 2.5%

α spending

- The set of α_k satisfy
  - Pr(any type 1 error) = 2.5% (or other needed overall alpha)

- We often refer to the "α spend" of a group sequential as
  - Pr(win at 1st interim | null) = a_1
  - Pr(win at 2nd interim | null) = a_2 (requires continuing at 1st interim)
  - …
  - Pr(win at final analysis | null) = a_K (requires continuing to end)

- Pr(win | null) = $a_1 + a_2 + \ldots + a_K = 0.025$ (or other desired value)

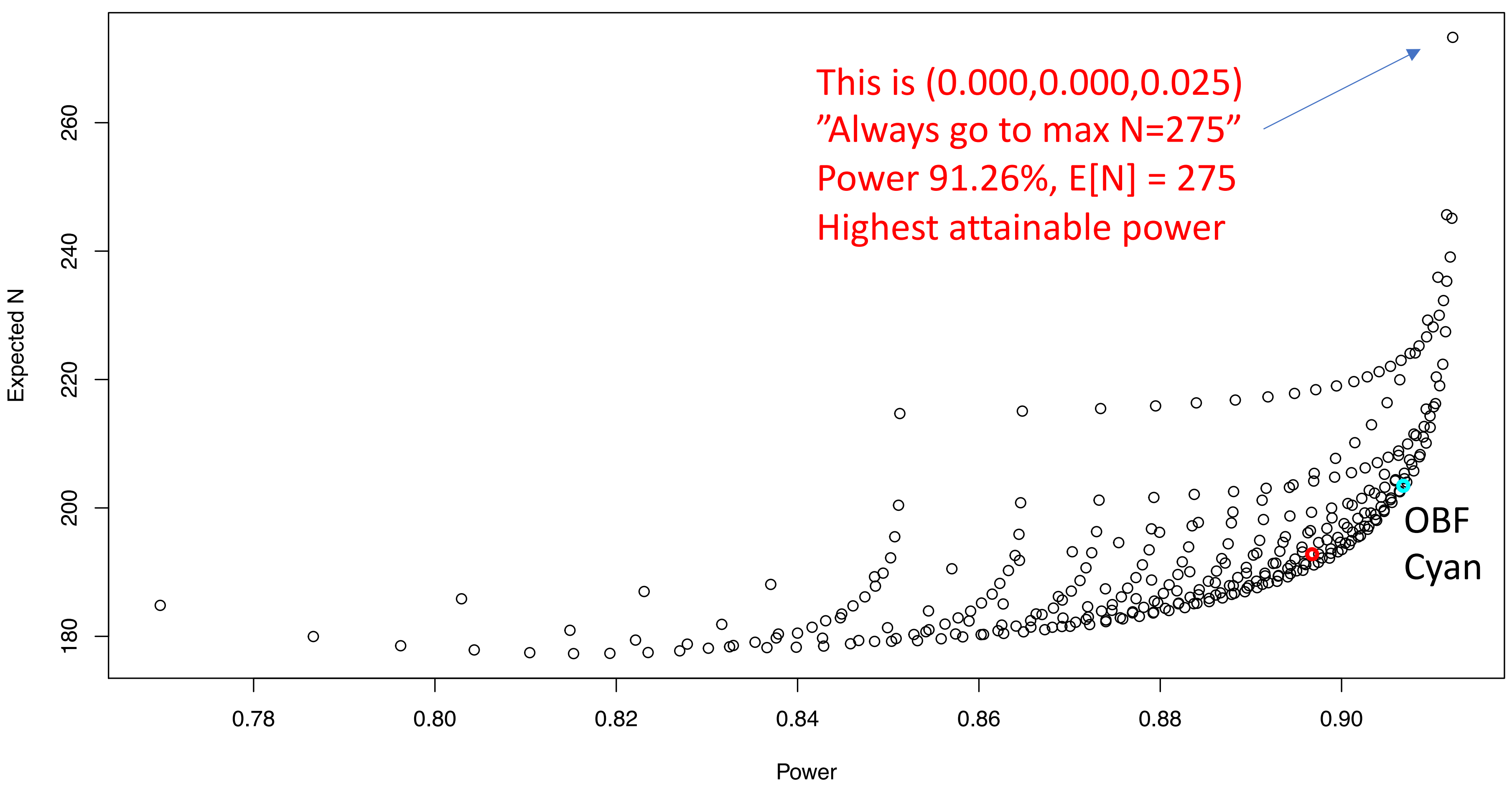
- Note α_k is not equal to a_k (the interims are correlated)
  - Given all N_k and a_k, can solve for α_k
  - Really only need n_k/n_K = information fractions (% of maximal size)

How to pick a_k?

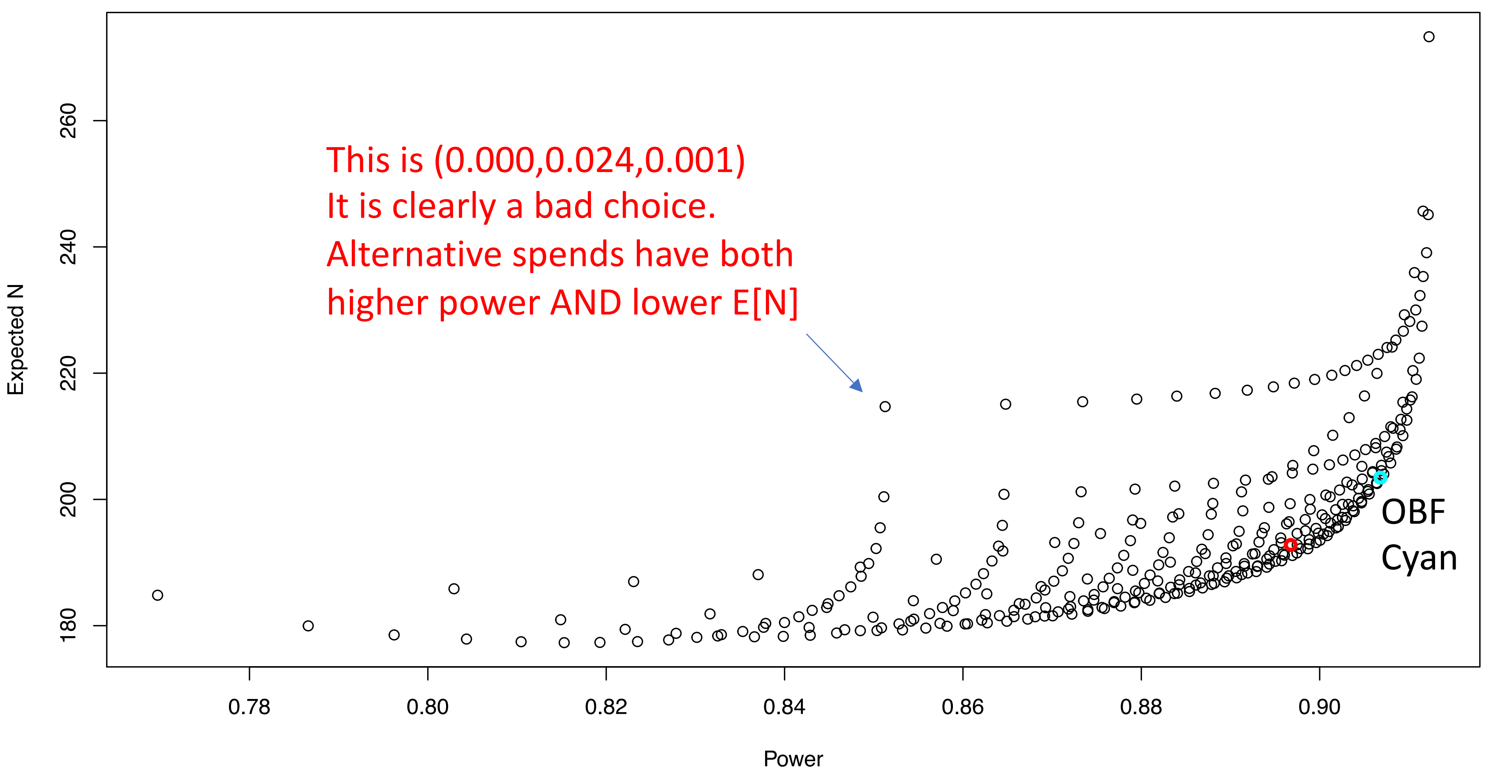
- Difference choices of a_k trade off sample size and power
  - some choices minimize sample size, others maximize power
  - some are just bad
- Let’s search ”all” possible a_k sequences for a specific trial
  - analyses at N=125, 200, 275
  - consider a grid of a_k sequences
    - (0.000,0.000,0.025), (0.000,0.001,0.024), (0.000,0.002,0.023), etc.
  - (0.000,0.000,0.025) is equivalent to always going to N=275
    - this had 91.26% power under hypothesized effectiveness.

- For each sequence, solve for α_k
  - find power and expected sample size for the trial

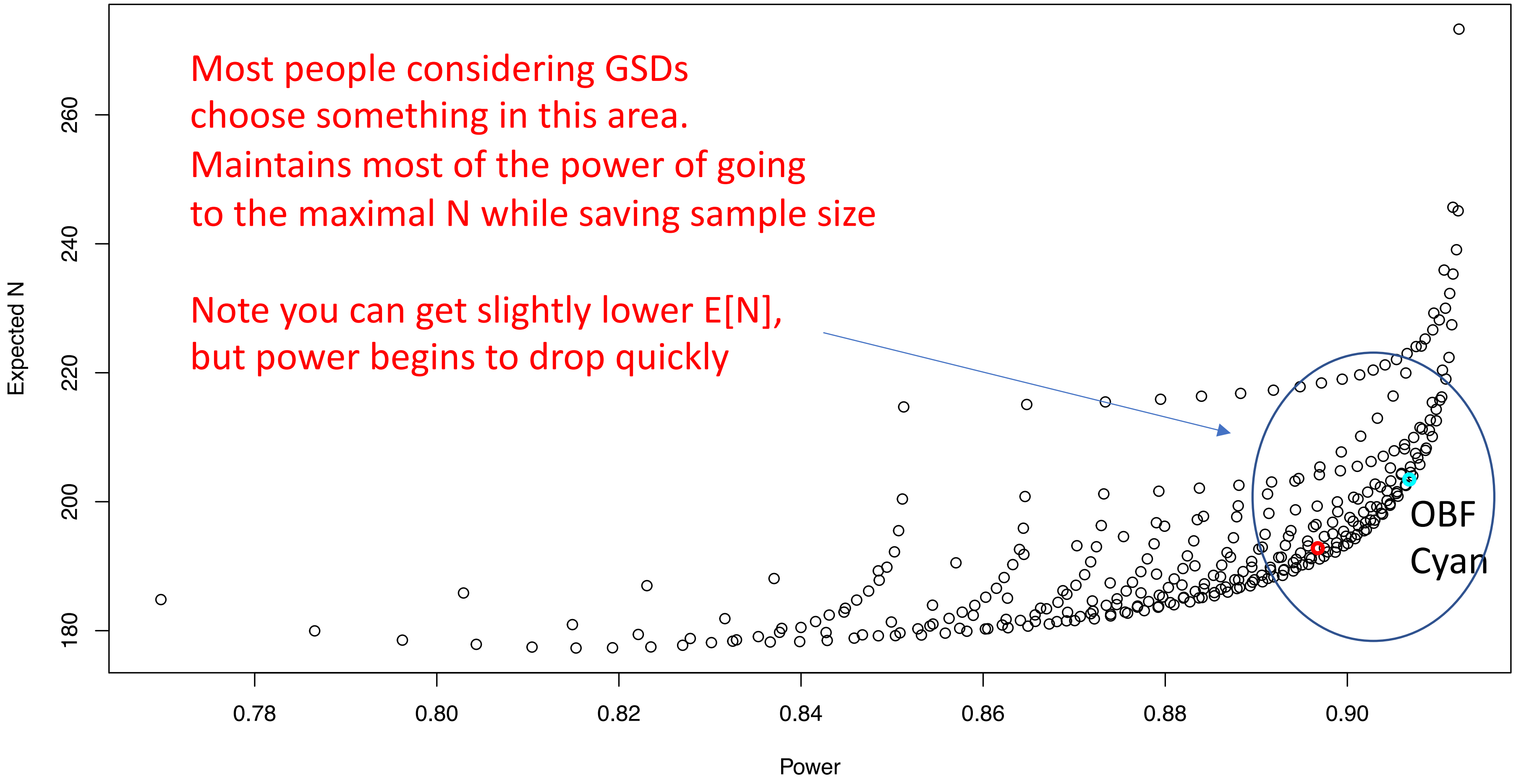
All possible trials in our grid



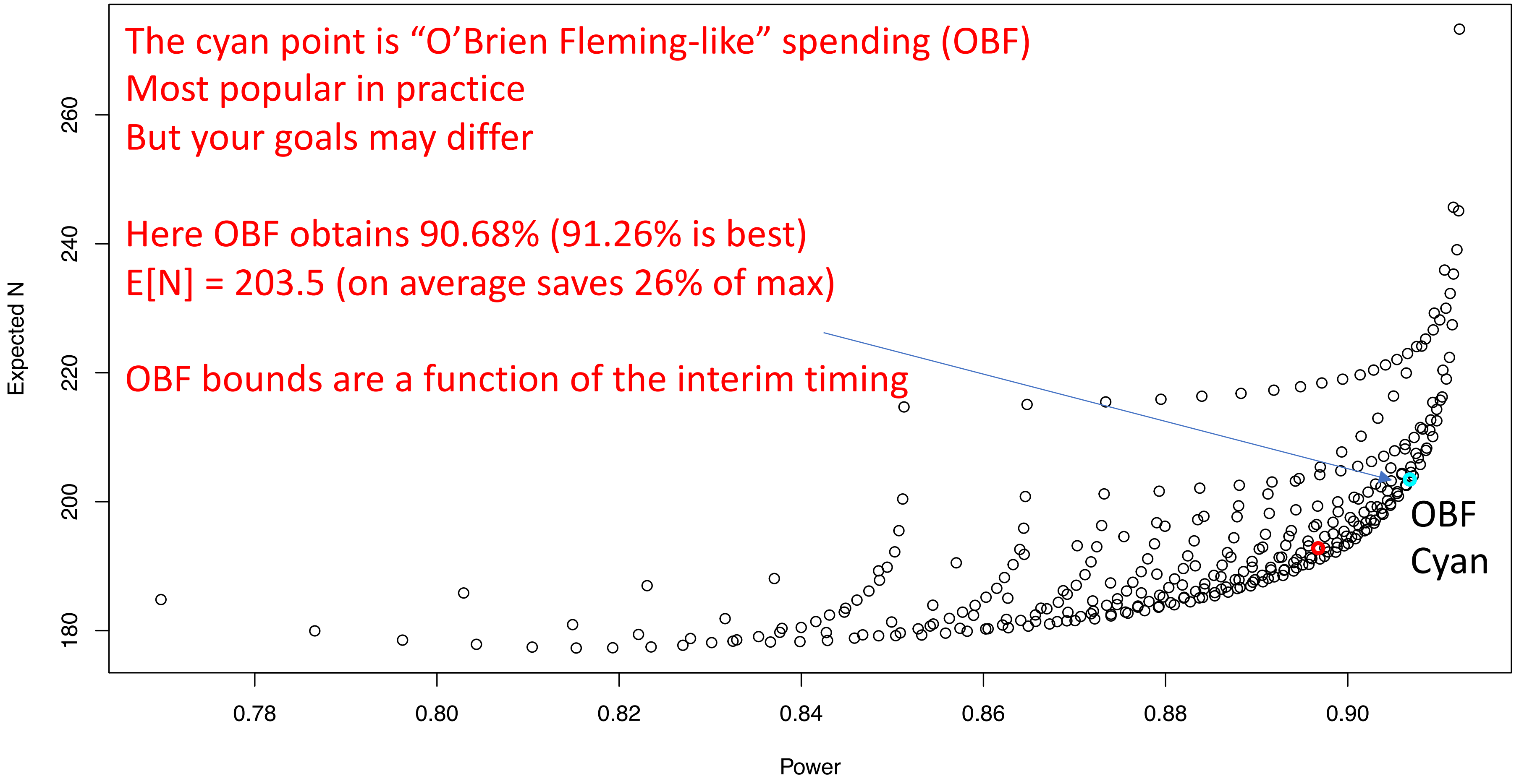
All possible trials in our grid



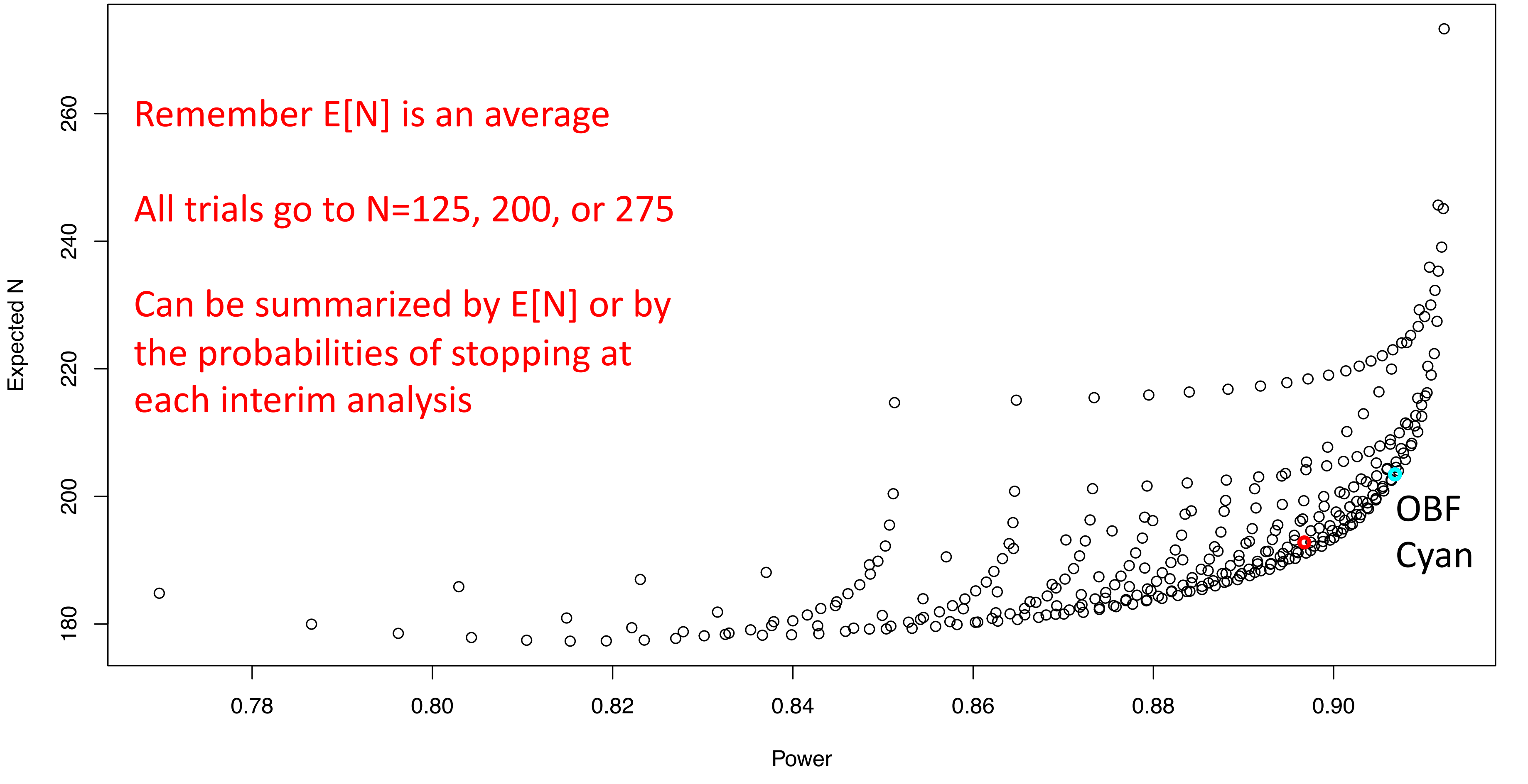
All possible trials in our grid



All possible trials in our grid



All possible trials in our grid



OBF like thresholds in R

```
library(gsDesign)
## Function to compute boundaries from Kim-DeMets spending function
getThresholds = function(looks, parameter, alpha = 0.025) {
  #relies on library(gsDesign)
  #Example
  #getThresholds(looks = seq(90, 210, 30), parameter = 3)
  #3 emulates OBF

  numlooks = length(looks)
  nmax = looks[numlooks]
  x1 = gsDesign(k = numlooks,
              timing = looks/nmax,
              test.type = 1,
              sfu = sfPower, sfupar = parameter,
              alpha = alpha)
  1-pnorm(x1$upper$bound)
}
```

```
> getThresholds(c(125,200,275),3,alpha=0.025)
[1] 0.002347859 0.008556151 0.021555638
```

Win if p1 (N=125) < 0.002347859, OR

Win if p2 (N=200) < 0.008556151, OR

Win if p3 (N=275) < 0.021555638

Choosing Interim Timing

- We arbitrarily chose N=125, 200, 275
- Are there better interim timings?

- First interim timing is extremely important
  - Sets smallest possible trial size, and thus caps efficiency
  - Need a “sufficient” minimal N (safety, secondary endpoints, etc.)

- Generally speaking
  - More interims is always statistically valuable (higher power, lower E[N])
  - Diminishing returns with high numbers of interims
  - Interims do have an operational cost

- We often vary first interim timing, consider lots of interims, and then remove interims as we refine the design if their operational costs exceed their benefits

A complete example trial

- Investigating a novel treatment
  - Dichotomous endpoint (response is good)
  - Anticipate control response rate 30% (null)
  - We hope our novel treatment has a 50% response rate (alternative)

- We could run a fixed N=200 (100 per arm) trial
  - one sided type 1 error = 2.5%, power = 83.3%

- Design as a group sequential, first interim at N=100
  - Analyses at 100, 120, 140, 160, 180, 200, 220 with OBF bounds
  - Maximal N=220 > 200 to maintain power

Group sequential version (with max N=220)

- Power increased to 85% (could have used N=210 or so?)
- Expected sample size N=156.4
- Compared to N=200 fixed, you essentially are playing a bet
  - 24.7% chance save 100, 10.9% chance save 80, …, 21.5% chance gain 20
  - the expected value of that bet is heavily in favor of the GSD.

Look	100	120	140	160	180	200	220
P-value required	0.0023	0.0031	0.0047	0.0069	0.0097	0.0134	0.0180
Pr(win)	0.2469	0.1086	0.1359	0.1094	0.0982	0.0858	0.0656
Pr(lose)	0	0	0	0	0	0	0.1495

Adding futility rules

- If the null is true, 97.5% of the time we go to N=220 and lose

- Berry tends to use predictive probabilities for futility
  - Compute probability trial will win from this point forward
  - If this probability if low, stop the trial for futility
    - avoid future costs with limited chance of benefit
    - how low depends on sponsor/funder goals
    - common choices 1%, 5%, 10%, 20%   (5% and 10% most common)

- Predictive probabilities incorporate uncertainty about the current treatment effect
  - Conditional power also possible (assumes treatment effect known)
  - Saville et al. The utility of Bayesian predictive probabilities for interim monitoring in clinical trials. Clin Trials 2014;11(4);485-493.
  - Saville, Detry, Viele. Conditional Power: How likely is trial success? JAMA 2023;329(6);508-509
  - Wendelberger, Lewis. Futility in Clinical Trials. JAMA 2023;330(8);764-765.

Example

- 140 patients into the trial (70 per arm)
  - 15/70 = 21% control, 19/70 = 27% treatment
  - current Z=0.79, p=0.2147
- What is the probability we win this trial?
  - We typically just compute Pr(meet success condition at N=220)
  - Pr(win at 220) approximates Pr(win at any future N)
- Need p<0.018 by N=220
  - backsolving, this requires approximately 15% observed effect
  - currently we have 6%, and we only have 80 patients to go
  - We need about a 32% effect on those 80 patients..doesn't feel likely

Computing the predictive probability

- 140 patients into the trial (70 per arm)
  - 15/70 = 21% control, 19/70 = 27% treatment
  - current Z=0.79, p=0.2147
- Place priors on the rates in each arm (Beta(0.5,0.5)?)
  - typically noninformative unless you have good prior data
- Posterior distributions
  - p_{ctrl} | data ~ Beta(15.5,55.5) p_{trmt} | data ~ Beta(19.5,51.5)
- Predictive distributions for the last 40 patients per arm
  - Y_{ctrl} ~ BetaBin (40,15.5,55.5) Y_{trmt} ~ BetaBin (40, 19.5, 51.5)
- Sidebar….a conditional power would assume p_{ctrl} and p_{trmt} are known to be their observed values

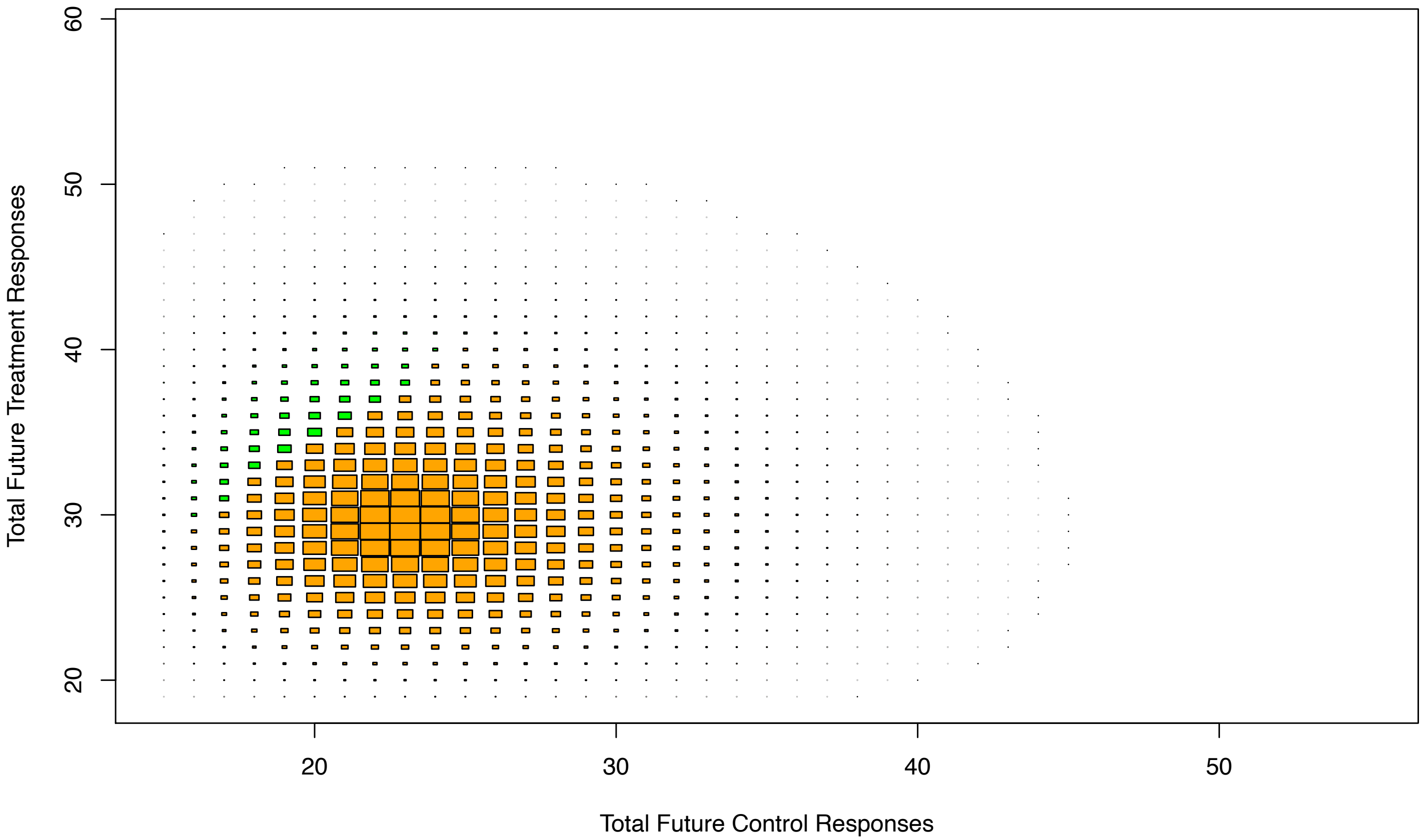
Graph showing predictive probability

Graph shows all combinations of future total control and treatment responses

Area of rectangle proportional to predictive probability of that combination

Green = successful ($p<0.018$)
Orange = not successful

Small probability of eventual success



Back to our example

- Let's add a rule to our example
  - Stop trial for futility if the predictive probability is less than 5%

- Managing a tradeoff between
  - Aggressive stopping saves sample size in the null
  - Can lose power in the alternative

- We often simulate 1%, 5%, 10%, 20% and discuss with the client
  - Choice can depend on client portfolio (opportunity costs)
  - Funders may be more aggressive than sponsors to declare futility

You can optimize a lot….

- Interim timing, alpha spending, futility thresholds all can significantly affect the value of a trial

- “Value” might be measured in terms of value to patients (getting a therapy to patients faster) or a sponsor may be interested in the financial value

- Properly valuing time, treatment effect, etc. is important
    - often simply approximated
    - packages available, QUOTES….

Operating Characteristics

Look	100	120	140	160	180	200	220
P-value required	0.0023	0.0031	0.0047	0.0069	0.0097	0.0134	0.0180
Pr(win)	0.221	0.115	0.127	0.129	0.103	0.085	0.052
Pr(lose)	0.031	0.015	0.013	0.016	0.016	0.023	0.054

Under alternative
20% treatment effect

power 83.3% (equals fixed trial)
expected N = 149.9
only 10.6% reach N=220

Look	100	120	140	160	180	200	220
P-value required	0.0023	0.0031	0.0047	0.0069	0.0097	0.0134	0.0180
Pr(win)	0.003	0.002	0.003	0.003	0.003	0.005	0.004
Pr(lose)	0.589	0.113	0.085	0.066	0.053	0.038	0.032

Under null
0% treatment effect

type 1 error < 2.5%
expected N = 123.1
80% of trials stop at
or before N=140

Real world impact

- Some trials involve nulls, some involve alternatives, some in between
  - we can imagine a distribution on the true effect trial to trial
- If that distribution were
  - 20% are our alternative (30% control, 50% treatment)
  - 80% are out null (30% control, 30% treatment)
- Our trials have equivalent power to running fixed trials
- Long run expected N per trial
  - (0.20 * 149.9) + (0.80 * 123.1) = 128.5
- Would allow us to fund over 50% more trials…
  - Note futility produces more of the savings than success…this is typical

Implementation in FACTS

- interactive outside slide deck

- Key items people like to change
  - Interim timing (Design/Interims)
    - will need to find revised thresholds in R or elsewhere
    - reenter thresholds (Design/Success and Futility Criteria)
    - revise 0.018 final threshold in predictive probability (Quantities of interest, Predictive probabilities)
    - max sample size can be changed in (Study/Study Info)
  - Futility threshold
    - Design/Success and Futility Criteria, easy to change at each interim

- Key performance metrics
  - Probability of early stops for success and futility shown in output
  - Expected sample sizes shown in output
  - Time Course for success and futility stopping shown in graph (exact numbers in the output files)

Are GSDs biased?

- It depends…on the plausibility of interim wins
- Note the overall conclusion of "superiority" is still fully type 1 error controlled, at easy is the point estimate
  - Viele, McGlothlin, Broglio. Interpretation of Trials that Stop Early. JAMA 2016;315(15);1646-1647

- It's always worth backsolving what effects are needed to win
  - Suppose at each interim we had a 30% observed control rate
  - What observed treatment rate is needed to win? Are these plausible?

Look	100	120	140	160	180	200	220
Needed p-value	0.0023	0.0031	0.0047	0.0069	0.0097	0.0134	0.0180
Needed observed treatment rate to win (ctrl=30%)	(31/50) 62.0%	(34/60) 56.6%	(38/70) 54.4%	(41/80) 51.2%	(44/90) 48.9%	(47/100) 47.0%	(50/110) 45.5%

Are GSDs biased?

- The first interim is always the most worrisome
  - Requires the most extreme results
  - Later interims less prone to bias because extreme results won earlier…
- N=100 wins with 30% control and 62% treatment (or better)
  - is a 32% treatment effect plausible?

Look	100	120	140	160	180	200	220
Needed p-value	0.0023	0.0031	0.0047	0.0069	0.0097	0.0134	0.0180
Needed observed treatment rate to win (ctrl=30%)	(31/50) 62.0%	(34/60) 56.6%	(38/70) 54.4%	(41/80) 51.2%	(44/90) 48.9%	(47/100) 47.0%	(50/110) 45.5%

Are GSDs biased?

- N=100 requires ~32% treatment effects
- Our alternative was 20% (30% vs 50%)
- If 20% was the maximum plausible effect, then observed 32% treatment effects will be biased high.
    - may want to consider removing this interim
    - bias corrections are possible, but Bayesian or frequentist they will involve an estimate that isn't that close to the data. May result in interpretation issues.

Bayesian view of GSD bias

- Suppose treatment effects from 0-40% were all equally likely
  - This will be my prior distribution
  - Assume control rate is 30% (can generalize)
  - Thus 32% treatment effect is plausible

- Suppose I observed 15/50 (30%) ctrl, 31/50 (62%) on trmt
  - Posterior mean treatment effect is 30.3%, slight reduction
  - Context dependent, but often worth reporting sooner rather than delaying effective treatment for 1-2% adjustment

- If effects from 0-20% were equally likely (32% impossible)
  - posterior mean treatment effect 17.1%, LARGE reduction
  - so different from observed data may create interpretation issues

Frequentist bias corrections

- Frequentist methods for bias correction exist as well

- Adaptive design guidance references
  - Jennison and Turnbull. Group sequential methods with applications to clinical trials. CRC Press.

- Our regulatory experience is primarily Bayesian, where the posterior distribution (posterior mean, credible intervals) is viewed as ”the answer”. We have less experience with frequentist corrections.

Practical Advice on Bias

- Backsolve the needed treatment effects to stop the trial early

- Ask as broadly as possible whether these effects would be believed and/or result in changing practice
  - If not, consider delaying the first interim
  - If so, potential biases are far more limited. Slight corrections from a prior distribution are sensible

- Sidebar….none of these biases are from “stopping early”
- The biases occur because you are requiring very small p-values with small sample sizes. A fixed trial (no early stopping) with the same requirements would produce the same biases.

Delayed Endpoints and Goldilocks trials

- Group sequential designs implicitly assume a “quick” endpoint
- If you have an interim at N=100
    - if a few patients are incomplete that is often minor
    - if 50 patients are incomplete…that is a very different issue.

- Number incomplete at each interim is a function of
    - endpoint time
    - accrual rate
    - e.g. with a 6 month endpoint and enrolling 25 patients a month, expect 150 patients incomplete at each interim

- Incomplete patients may supply information (e.g. early visits)

Why is a lot of incomplete data a problem?

- Interpretation
  - If we stop a trial at an interim analysis, we have two data sets to consider, the interim dataset and the full followup
  - With lots of incomplete data at interim, these may be materially different. Even carefully defining “the primary analysis”, differences can create “review issues”.

- Efficiency
  - Incomplete patients provide less information than complete patients
  - But if we wait for information, it’s hard to stop a trial meaningfully early (e.g. the trial may be nearly enrolled before many patients reach their final endpoint).
  - In many trials, we may be able to meaningfully use partial information from incomplete patients

Goldilocks Strategy

- At each interim, compute two predictive probabilities
  - Pr(win trial | stop now and followup) = PPn
    - includes uncertainty in followup
  - Pr(win trial | continue to max N) = PPmax
    - includes uncertainty in followup and future patients
    - similar/identical to our prior futility calculations

- These predictive probabilities may include a longitudinal model predicting final outcomes from available patient information
  - for example, a patient who has not experienced a major adverse event by 3 months may be unlikely to have an AE before 6 months.

Goldilocks Strategy using PPn and PPmax

- Stopping accrual for anticipated success
  - Stop if PPn > Sn (Sn can vary by interim, but often doesn't)
  - Final analysis usually occurs at full followup
    - sometimes final analysis may occur at interim, but often regulators don't want to make decisions on a dataset with large amounts of incomplete data
    - additionally, may lack information on secondary and other endpoints at the interim for operational reasons
  - At full followup, declare success if p-value<Bn
    - Bn may differ by interim, often the same for each n
    - in confirmatory trials, Sn and Bn must be selected to maintain type 1 error control, typically demonstrated by simulation
    - Can also declare final success based on a posterior probability

Goldilocks Strategy using PPn and PPmax

- Stopping trial for futility
  - Stop for futility if PPmax < Fn
  - e.g. limited chance of trial success, even if we enrolled until the end.
    - this often approximates the chance of ANY success, which is harder to compute

- Continue trial if neither PPn > Sn or PPmax < Fn
  - stop at prespecified maximal sample size if reached
  - again, trial success if posterior probability at final > Bn

- Broglio, Connor, Berry. Not too big, not too small: a goldilocks approach to sample size selection. J Biopharm Stat 2014;24(3);685-705.

Example

- Single arm trial in oncology
  - dichotomous endpoint (patient response)
  - Need to show superiority to an OPC rate of p=0.20
  - Hoped for improvement to p=0.35 response rate
  - Accrual 1 patient/week, endpoint is at 17 weeks

- Exact binomial test with fixed sample size N=100
  - requires 29/100 to obtain significant (2.5% type 1 error)
  - trial has 91.5% power when p=0.35

- Can we make this smaller?
  - Or equivalently suppose we felt response rates from 35-50% were plausible. 50% response rates naturally require smaller N.
  - Do not wish to have fewer than 50 patients in the trial

Goldilocks strategy

- At 1 patient/week and a 17 week endpoint
  - expect 17 incomplete patients at any given
  - decent fraction of our total data, suggests Goldilocks strategy

- Conduct interims when 50, 60, 70, 80, 90 patients enrolled
  - Compute PPn and PPmax
  - Stop for success if PPn > 90%
  - Stop for futility if PPmax < 5%
  - After full followup, need p<0.03 to win

Operating Characteristics

- Recall a fixed trial N=100
  - always goes to 100, so E[N]=100
  - power 91.5%

Scenario	p=0.20 (null)	p=0.30	p=0.35	p=0.40
Pr(trial success)	0.0206 type 1 error	0.5977	0.8849 power	0.9805
Expected N	67.8	82.1	73.1	62.6

Longitudinal Information

- Beyond current scope of talk

- Often early visits convey information
  - A knee device patient who is successful at 6 months is likely to remain successful at 1 year, 2 years, etc.
  - Often a failure at an early endpoint implies failure at the final endpoint (for example presence of adverse event)

- FACTS supports longitudinal modeling
  - Beta Binomial imputation
  - predict final visit from each interim visit
  - Continuous variants
  - Pro tip – to assess whether longitudinal modeling will be helpful, reduce the endpoint time and see whether performance increases

FACTS implementation

- Interactive demo outside slide deck

- Key Items to change
  - Interim Schedule (Design/Interims)
  - Success and Futility thresholds (Design/Success and Futility)
  - Final success condition (Quantities of Interest)
  - Accrual rate (Execution/Accrual)

- Key performance metrics
  - Type 1 error rate (Simulation output)
  - Power and expected sample size (Simulation output)
  - Time course of stopping (Simulation output graph)

Regulatory/Operational Issues

- Group sequentials and Goldilocks well accepted by regulators
  - Approvals for both methods
  - Many goldilocks in FDA/CDRH

- Common issues in review
  - justifying number of interims (many interims are ok, but you need to show meaningful increase in performance)
  - binding vs non-binding futility
    - You may need to show the trial is type 1 error controlled even if futility is turned off (guidance allows either, our experience is that non-binding is preferred)
  - Concerns about operational bias

Operational Bias

- Operational bias refers to the effect of data “leakage” on the conduct of the trial
  - If you put out a press release saying “the trial is continuing after interim 2”, does that provide external people information.

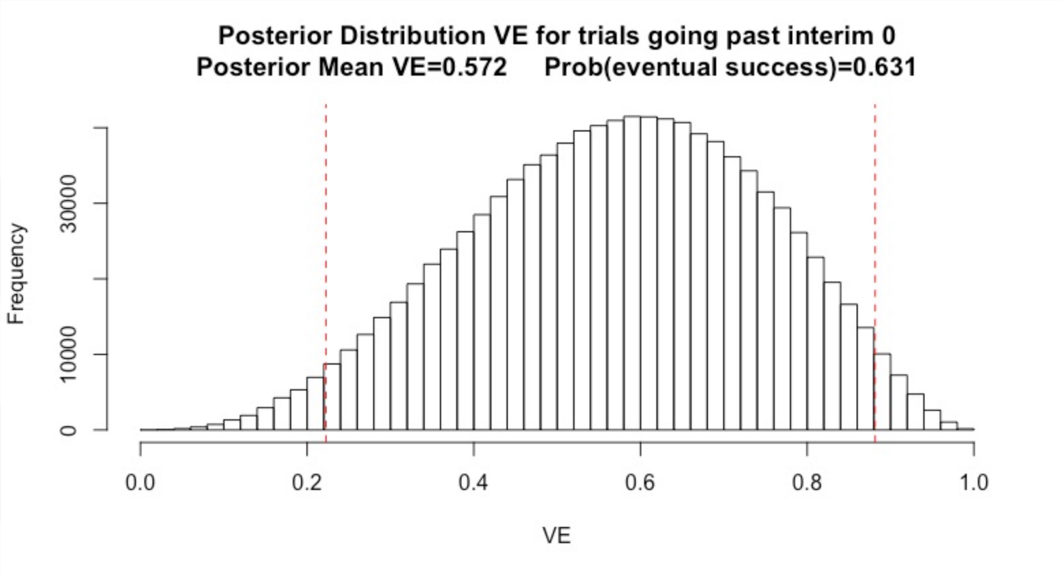
- We are often asked by venture capitalists “here is the publicly available information, is the trial going to win?”

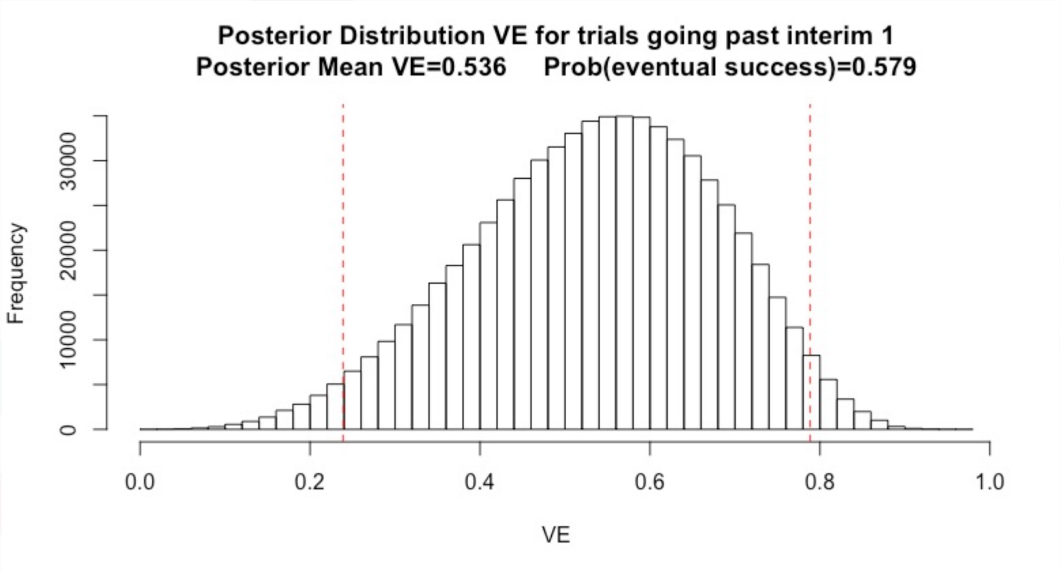
- Generally speaking, group sequentials leak limited information
  - Knowing the trial is continuing doesn’t meaningfully change the predicted probability of success
  - You do eliminate “extreme” possibilities from consideration

Pfizer Vaccine trial example

- Pfizer had interim analyses based on events
  - Number of trial participants diagnosed with COVID-19

- https://twitter.com/KertViele/status/1307463136736354308





Thank you

- Thank You for attending
- Link to Recording will be sent out tomorrow
- Slides will be available via our website at the end of the series
- Any questions please contact us:
  - tom@berryconsulants.com
  - kert@berryconsultants.com
  - facts@berryconsultants.com
- If you would like a demo and/or a free evaluation copy of FACTS
- Berry regularly produces blogs and social media posts on adaptive designs
  - @KertViele, Kert Viele on LinkedIn