Evaluating an Adaptive Clinical Trial with Quantitative Endpoints, Sample Size Re-estimation, Sequential Monitoring for Efficacy, and Monitoring for Futility

By: Harrison Reeder and Kamrine Poels
Mentor: Dr. Kathryn Chaloner

Outline

- What exactly does that title mean?
  - Basic Clinical Trial design
  - Interim Monitoring for Efficacy
    - 3 schemes for interim monitoring for efficacy
  - Interim monitoring for futility
  - Adaptive sample size re-estimation
- Simulation Study of Design Performance
- Conclusion

Clinical Trial Design: The Basic Case

- The most basic element of clinical trial design is determining an adequate sample size

- Calculating sample size requires specifying:
  - approximate variance of outcomes
  - the desired Type I error rate
  - minimum clinically meaningful treatment effect
  - desired power to detect that effect

power.t.test()

Interim Monitoring for Efficacy

- Why use interim monitoring?
- Complications of interim monitoring

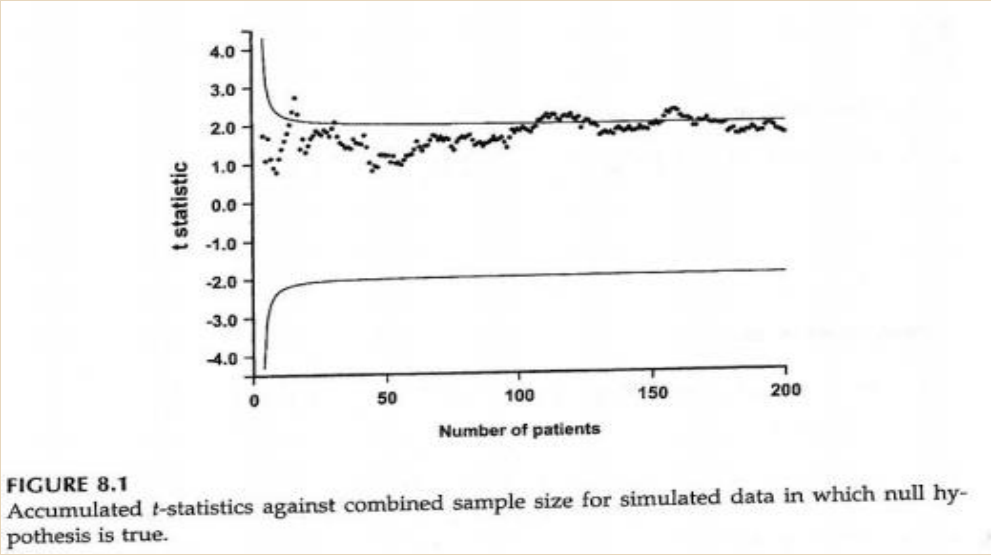


FIGURE 8.1
Accumulated *t*-statistics against combined sample size for simulated data in which null hypothesis is true.

Interim monitoring inflates Type I error

Solution: Change boundary of significance

Taken from *Introduction to Randomized Controlled Clinical Trials* by John Matthews

Schemes for Interim Efficacy Monitoring

- Pocock "constant" boundaries
  - sets constant p-value boundary to use at every monitoring point
  - Earlier rejection is easier, but final test is stringent
- O'Brien-Fleming boundaries
  - makes rejection harder at earlier points and easier as trial progresses
- Fleming-Harrington-O'Brien boundaries
  - middle-ground between above strategies

Boundary	First Interim	Second Interim	Third Interim	Final point
Pocock	0.0182	0.0182	0.0182	0.0182
O-F	0.00005	0.0039	0.0184	0.0412
F-H-O	0.0067	0.0083	0.0103	0.0403

Interim Monitoring for Futility

- Why monitor for futility?
- Conditional power
  - Estimates probability of having significant results given observed data and (design) assumptions
  - If probability is lower than a specified threshold, then trial is stopped

Adaptive Sample Size Recalculation

- Early estimate of response variance is difficult
- To account for difference between estimate and true value, this design uses observed estimated variance halfway into trial to re-estimate sample size
- Investigators can set a maximum sample size for each group

Research Question: How does our design perform?

- Using simulation, we compare the design to designs without the features described
  - We also compare the merits of the three interim monitoring schemes
- Values of interest:
  - Bias of final treatment effect estimate
  - True confidence of nominal 95% Confidence Interval
  - True Type I error
  - True power
  - Distribution of stopping points

Designing the Simulation

First Interim
- Sample Size is 9
- Check for efficacy

Second Interim
- Sample Size is 18
- Check for efficacy
- Check for futility
- Final sample size is recalculated

Third Interim
- Sample size is $\frac{Final+18}{2}$ if recalculated
- Without sample size recalculation, size is 28
- Check for efficacy

Final Point
- Sample size is *Final* $\leq$ 50 if recalculated
- Without sample size recalculation, size is 35
- Check for efficacy

Motivating Study: Effect of Sleeping Drug in Adolescents and Young Adults with Autism Spectrum Disorder

Design assumptions:

- Mean treatment effect: 32 minutes
- Response standard deviation: 36 minutes

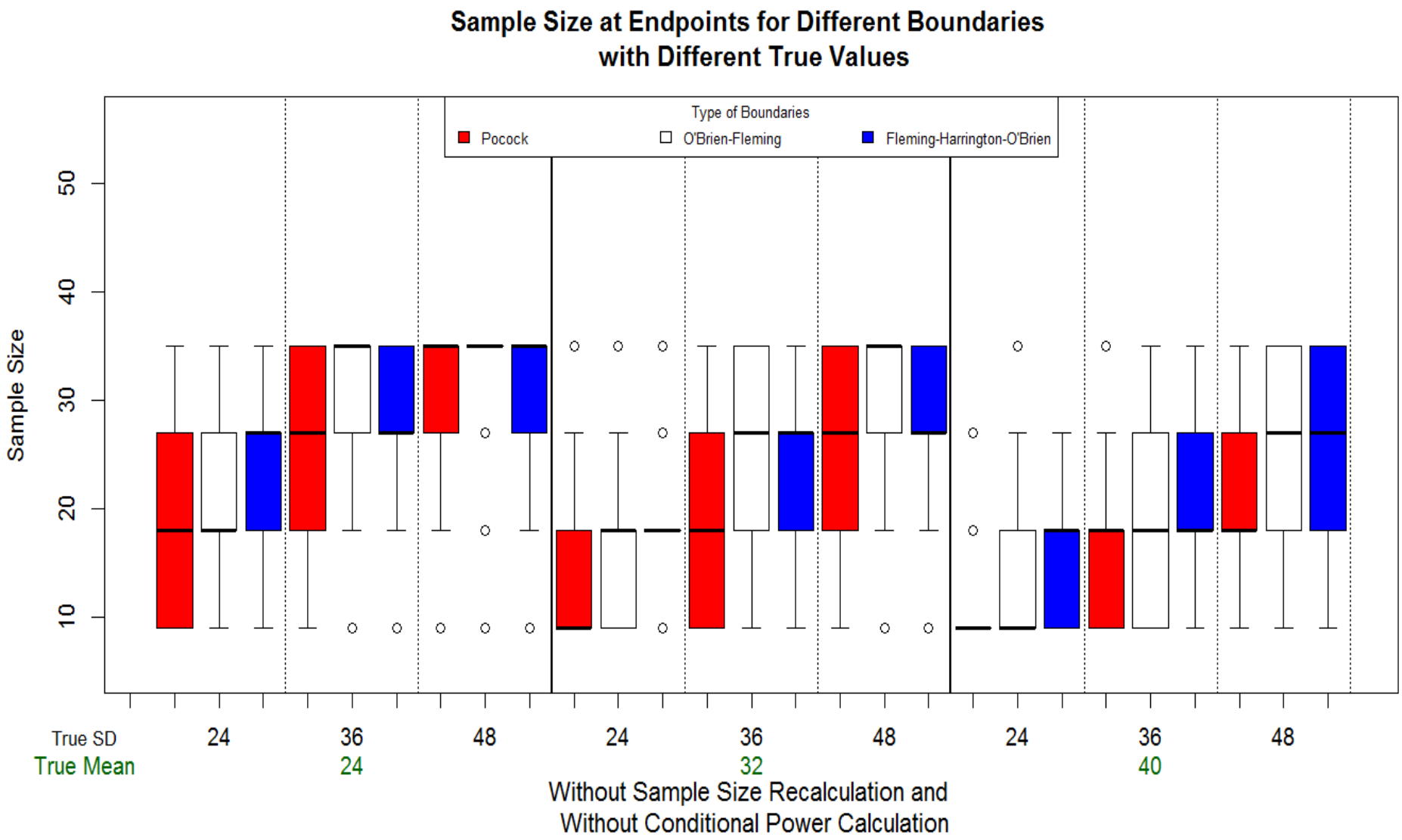
Simulation seed: 42
Conditional power seed: 123

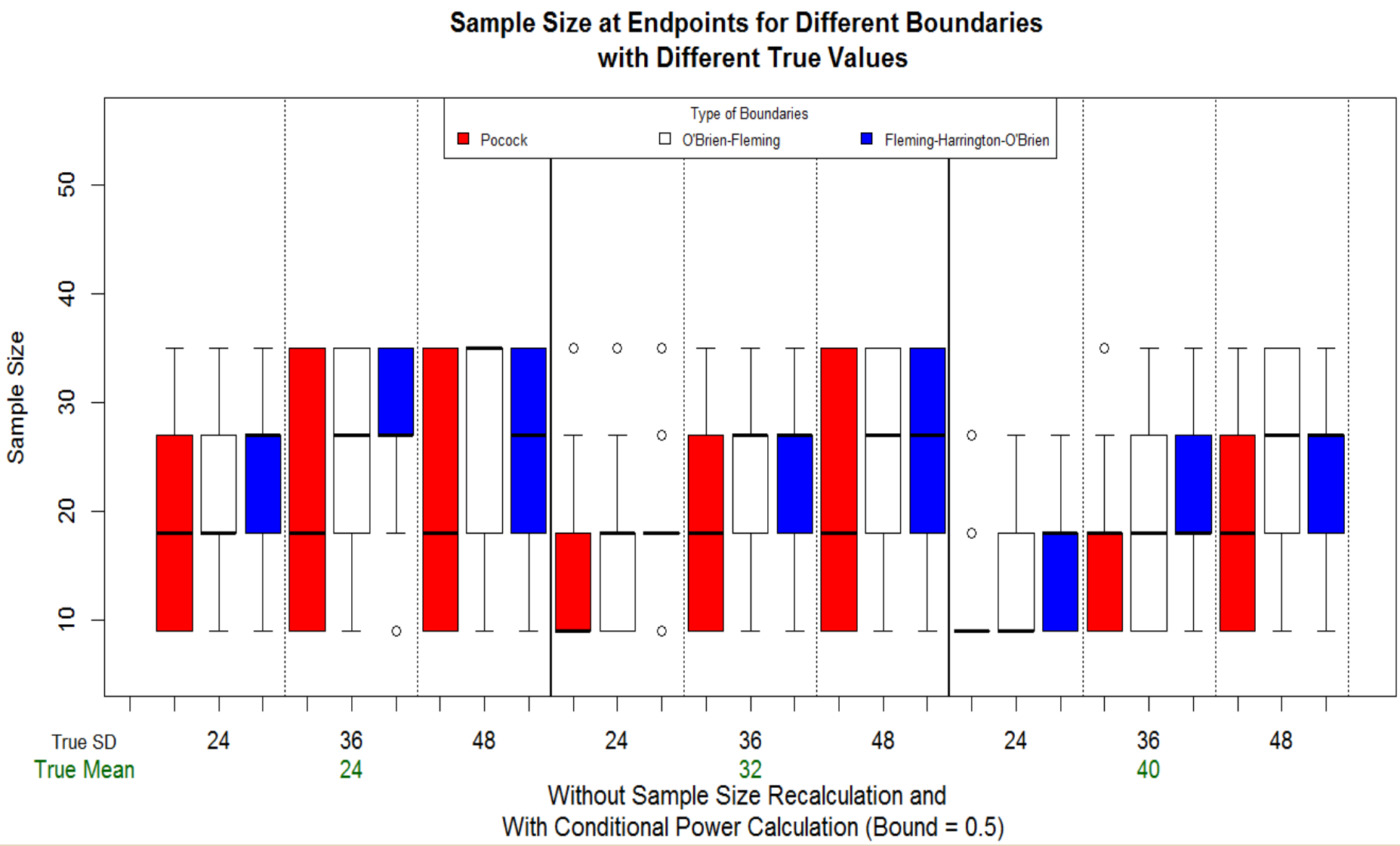
Effect of Interim Monitoring for Efficacy (Without Sample Size Re-estimation or Futility Monitoring)

- Ending sample size $\leq$ 35 per group because we can stop at earlier interim points when results are significant
- Bias of estimated treatment effect is positive (overestimates by ~10% on average)

Effects of Interim Monitoring for Futility
(Without Sample Size Re-estimation)

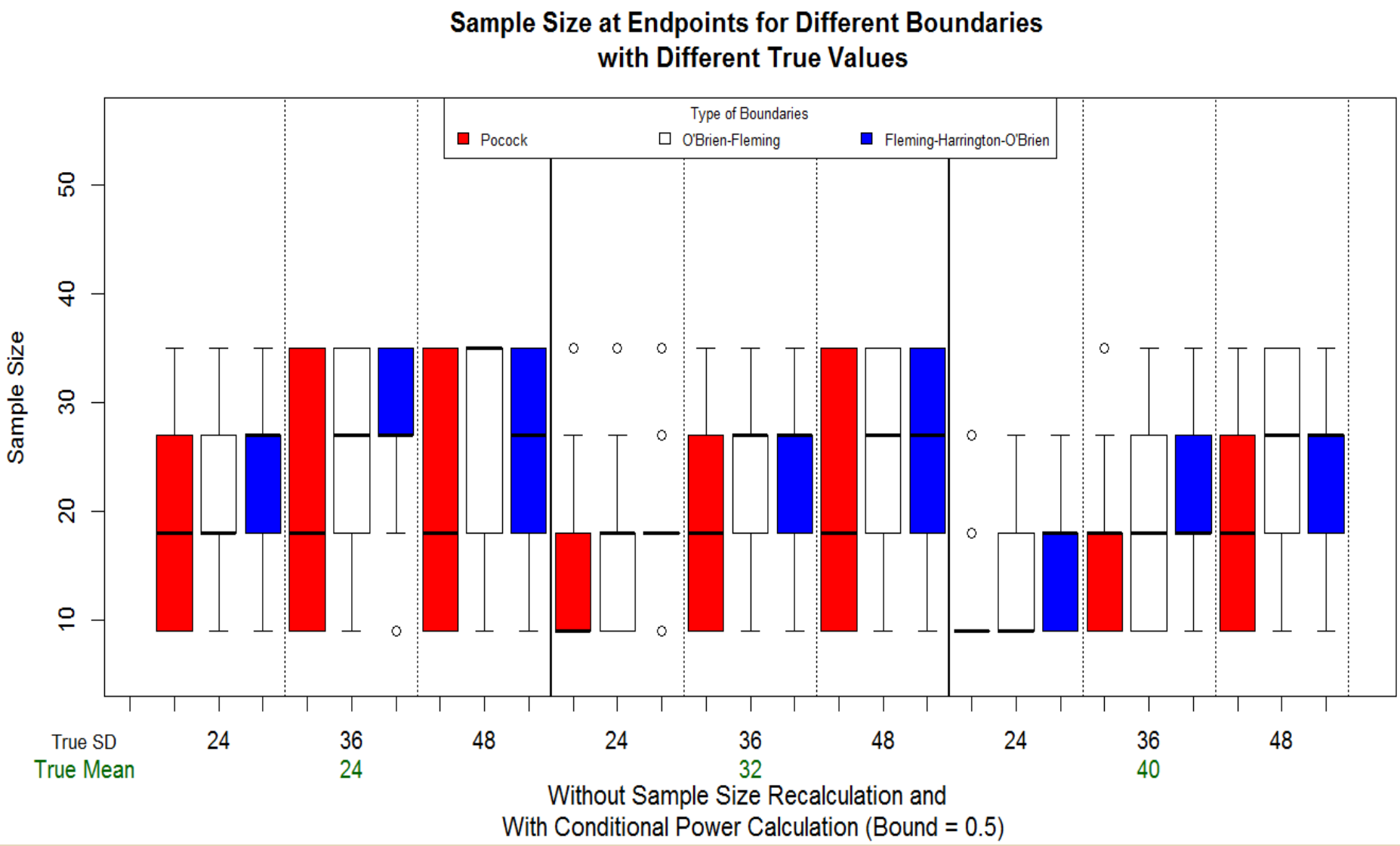
- Large drop in true Type I error from ~0.05 to ~0.01 (more opportunities to stop an ineffective trial from following through to the end and having significance by chance)
- Effects mediated by conditional power bound
- Smaller stopping point sample size when response variance is larger than expected
  - Chance of stopping early for futility, even if alternative is true, explains a slight drop in true power

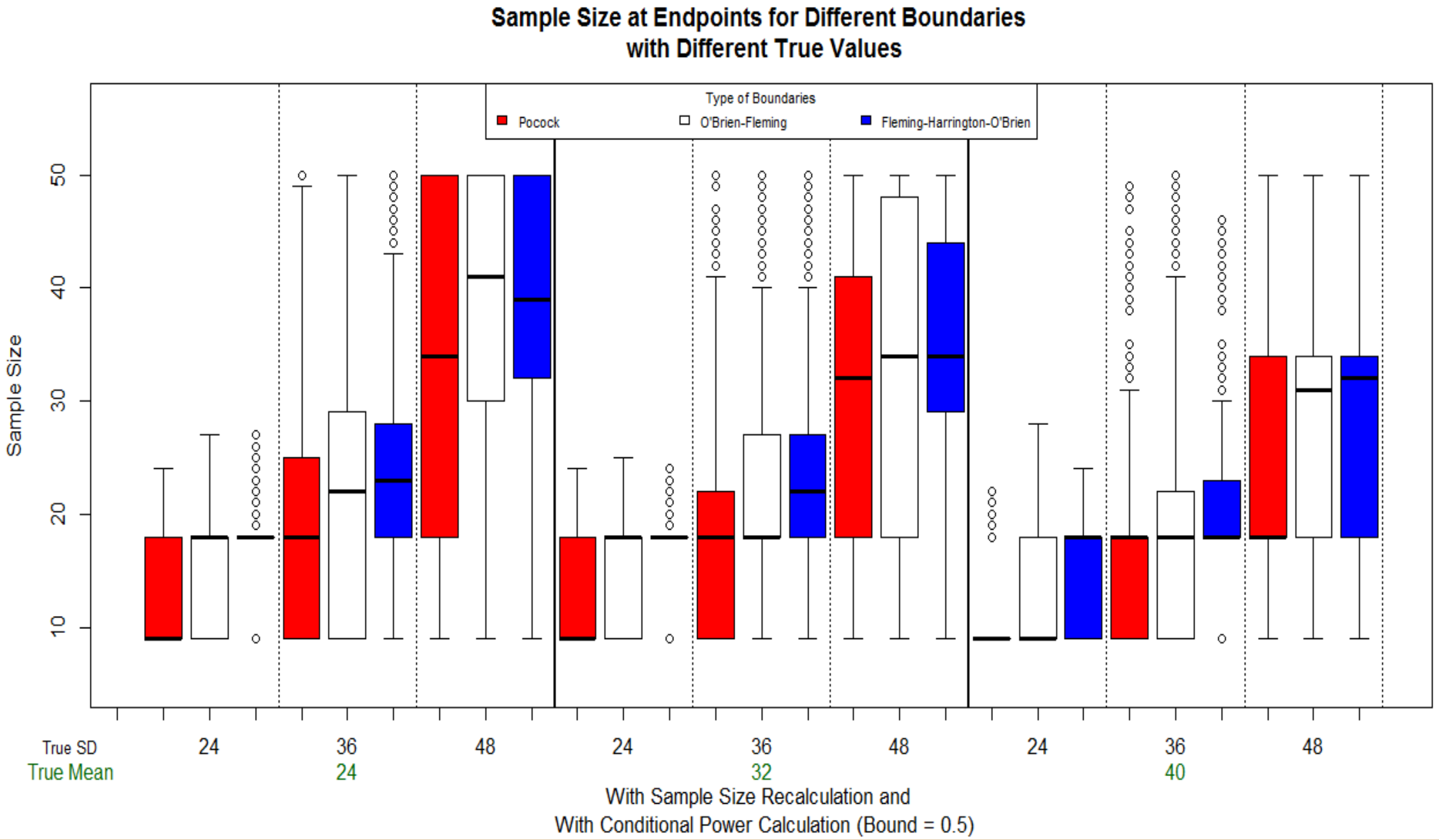

Sample Size at Endpoints for Different Boundaries
with Different True Values
Type of Boundaries
Pocock
O'Brien-Fleming
Fleming-Harrington-O'Brien
Sample Size
10
20
30
40
50
True SD
24
36
48
24
36
48
24
36
48
True Mean
24
32
40
Without Sample Size Recalculation and
Without Conditional Power Calculation


Sample Size at Endpoints for Different Boundaries
with Different True Values
Type of Boundaries
Pocock
O'Brien-Fleming
Fleming-Harrington-O'Brien
Sample Size
10
20
30
40
50
True SD
24
36
48
24
36
48
24
36
48
True Mean
24
32
40
Without Sample Size Recalculation and
With Conditional Power Calculation (Bound = 0.5)

Effects of Sample Size Recalculation

- If the design variance is greater than or equal to the true variance, recalculation tends to decrease the ending sample size
  - Likewise, underestimated variance leads to a larger required sample size
- Power follows a similar trend


Sample Size at Endpoints for Different Boundaries
with Different True Values
Type of Boundaries
Pocock
O'Brien-Fleming
Fleming-Harrington-O'Brien
Sample Size
10
20
30
40
50
True SD
24
36
48
24
36
48
24
36
48
True Mean
24
32
40
Without Sample Size Recalculation and
With Conditional Power Calculation (Bound = 0.5)


Sample Size at Endpoints for Different Boundaries
with Different True Values
Type of Boundaries
Pocock
O'Brien-Fleming
Fleming-Harrington-O'Brien
Sample Size
10
20
30
40
50
True SD
24
36
48
24
36
48
24
36
48
True Mean
24
32
40
With Sample Size Recalculation and
With Conditional Power Calculation (Bound = 0.5)

Comparison of Boundary Types

- Pocock
  - Highest Type I error
  - Highest bias
  - Lowest power
  - Smallest sample size (i.e., best chance of finding efficacy early)
- O'Brien-Fleming and Fleming-Harrington-O'Brien
  - Similar results across measures and assumptions
    - O'Brien-Fleming boundary is more commonly used

Overall Evaluation of Our Design

These characteristics show the design's potential value in Phase II trials:

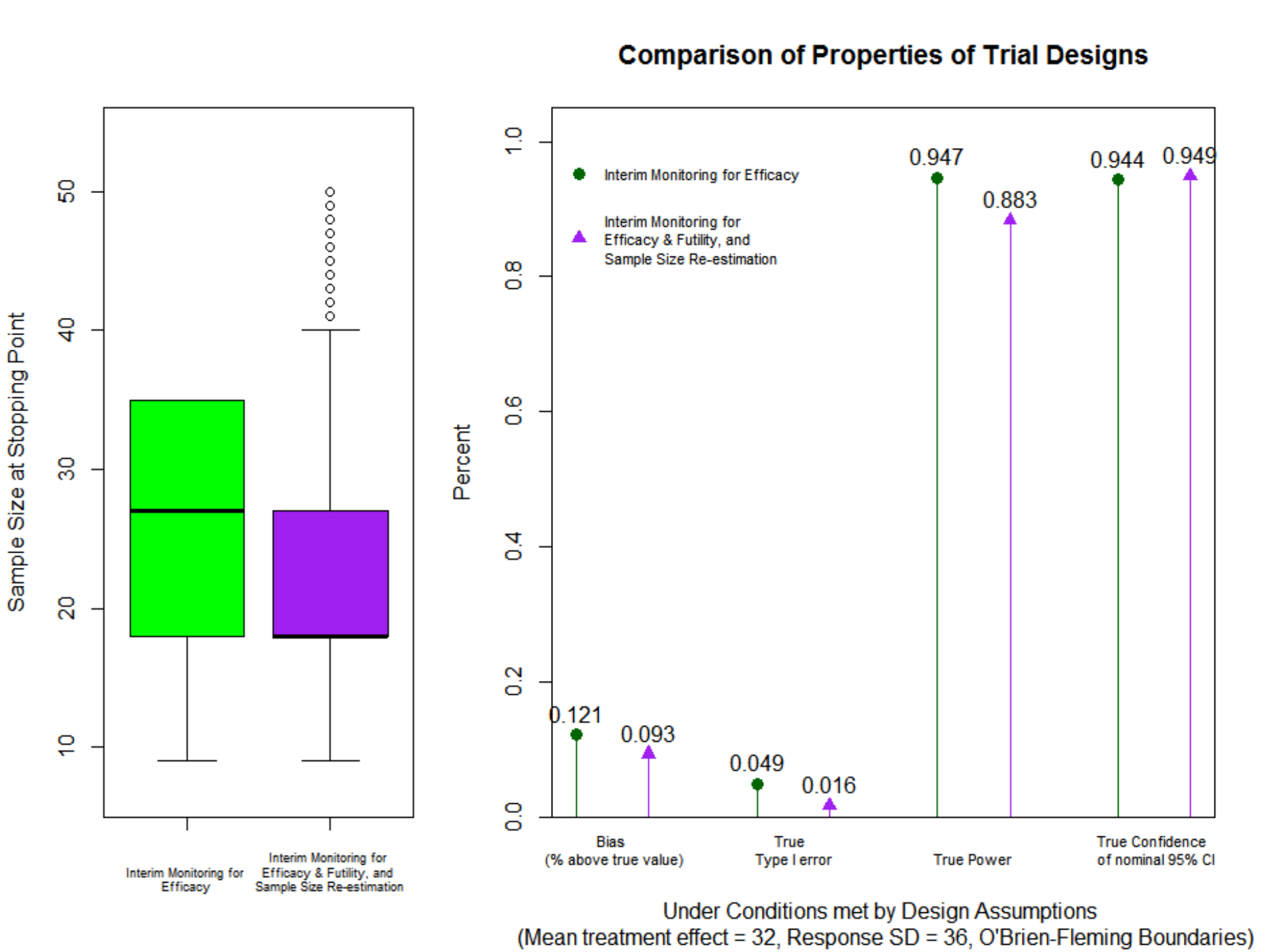
- Minimizes Type I error rate
- Maintains power when variance estimate is too low
- May decrease sample size required to reach a conclusion

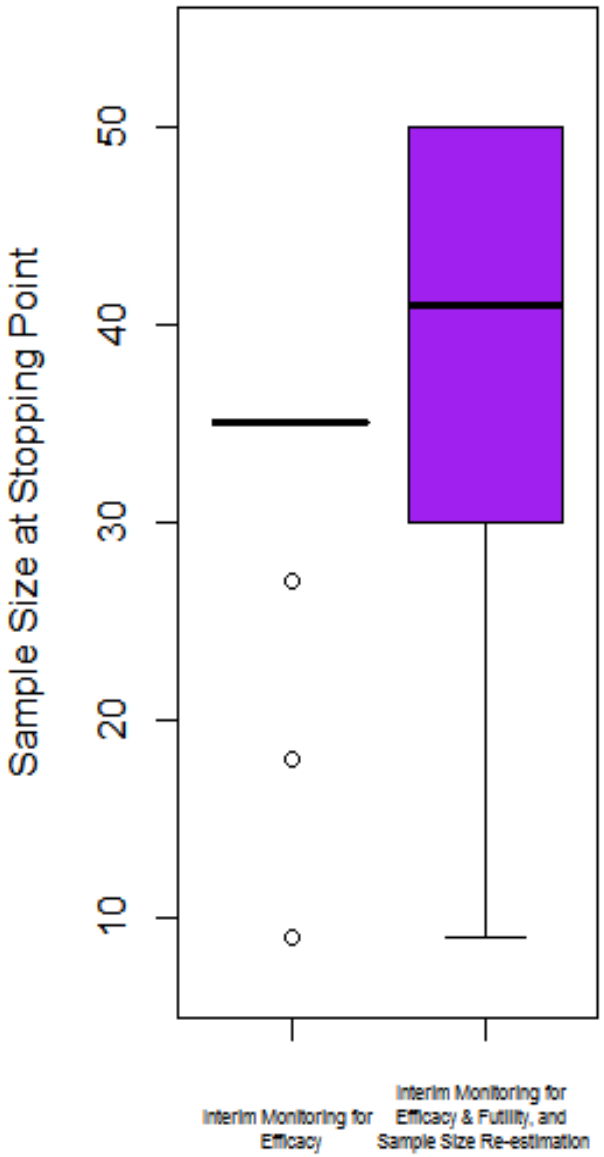
Limitations:

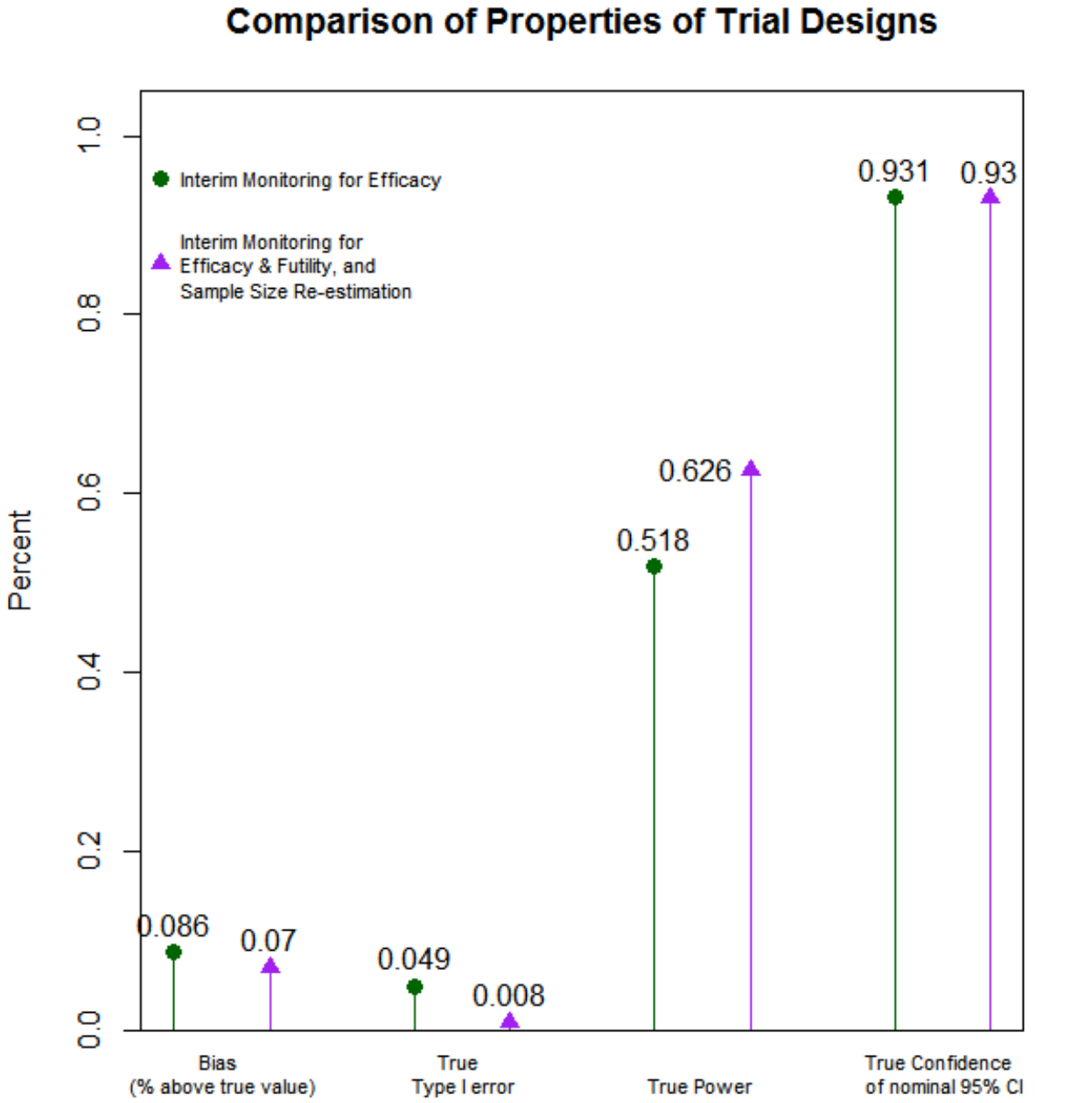
- Sample size re-estimation potentially increases cost
- Gives biased estimate of treatment effect

How Does Our Design Compare to Interim Monitoring for Efficacy Alone?

- If assumptions are accurate, with our design:
  - Median ending sample size is smaller
  - Power is slightly lower, but comparable
  - Type I error rate is lower (important for Phase 2 trials)
- If assumptions are inaccurate (overestimated effect size and underestimated variance):
  - Ending sample size tends to be larger (more expensive)
  - Power is higher (though overall both are much lower)
  - Type I error rate is lower


Sample Size at Stopping Point
10
20
30
40
50
Interim Monitoring for Efficacy
Interim Monitoring for Efficacy & Futility, and Sample Size Re-estimation
Comparison of Properties of Trial Designs
Percent
0.0
0.2
0.4
0.6
0.8
1.0
Interim Monitoring for Efficacy
Interim Monitoring for Efficacy & Futility, and Sample Size Re-estimation
0.121
0.093
0.049
0.016
0.947
0.883
0.944
0.949
Bias (% above true value)
True Type I error
True Power
True Confidence of nominal 95% CI
Under Conditions met by Design Assumptions
(Mean treatment effect = 32, Response SD = 36, O'Brien-Fleming Boundaries)


Sample Size at Stopping Point
10
20
30
40
50
Interim Monitoring for Efficacy
Interim Monitoring for Efficacy & Futility, and Sample Size Re-estimation


Comparison of Properties of Trial Designs
Percent
0.0
0.2
0.4
0.6
0.8
1.0
Interim Monitoring for Efficacy
Interim Monitoring for Efficacy & Futility, and Sample Size Re-estimation
0.086
0.07
0.049
0.008
0.518
0.626
0.931
0.93
Bias (% above true value)
True Type I error
True Power
True Confidence of nominal 95% CI
Under Conditions Overestimating Effect and Underestimating Response Variance
(Mean treatment effect = 24, Response SD = 48, O'Brien-Fleming Boundaries)

Conclusion:
"Is our design better for the motivating study?"

Yes!

- Minimizing Type I errors is important in Phase II trials, which is achieved in our design
- Treatment effect and response variance are not easily estimated in the motivating study
  - Our design's ability to maintain power and keep error rates low even with inaccurate design assumptions is beneficial

Limitation:

- Potential for higher re-estimated sample size may increase cost of trial