2024 Summer Institute In Statistics for Clinical & Epidemiological Research

Module 3:

Design, Conduct, and Analysis of Randomized Clinical Trials with Time to Event Primary Endpoints

Lecture 20:

Group Sequential Boundaries

Scott S. Emerson, M.D., Ph.D.
Professor Emeritus of Biostatistics
University of Washington

1

Without Loss of Generality

- Our ultimate interest is in comparing
  - Fixed sample tests
  - Group sequential tests
  - Other adaptive strategies
- We will thus further restrict attention to a one-sample setting in which
  - $V = 1$
  - Test of a one-sided alternative ($\theta_+ > \theta_0$)
    - Upper Alternative: H_+: $\theta \geq \theta_+ = 3.92$ (superiority)
    - Null: H_0: $\theta \leq \theta_0 = 0$ (equivalence, inferiority)

2

2

Fixed Sample Test

- Sample size N = 1 provides
  - Type 1 error of 0.025
  - Power of 0.975 to detect the alternative of 3.92
  - At the final analysis, an observed estimate (or Z statistic) of 1.96 will be statistically significant
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	1.00
1.96	0.500	1.00
2.80	0.800	1.00
3.24	0.900	1.00
3.92	0.975	1.00

3

3

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
  - Often chosen according to some boundary shape function
    - O'Brien-Fleming, Pocock, Triangular, …
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue

4

4

Stopping Boundary Scales

- Boundary scales (1:1 transformations among these)
  - Z statistic
  - P value
    - Fixed sample (so wrong)
    - Computed under sequential sampling rule (so correct)
  - Error spending function
  - Estimates
    - MLE (biased due to stopping rule)
    - Adjusted for stopping rule
  - Conditional power
    - Computed under design alternative
    - Computed under current MLE
  - Predictive power
    - Computed under flat prior (possibly improper)

5

5

Exploring Group Sequential Designs

- Candidate designs
  - J = 2 equal spaced analyses; O'Brien-Fleming efficacy boundary
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, futility boundaries
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 equal spaced analyses; OBF efficacy, more efficient futility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 4 equal spaced analyses; OBF efficacy, more efficient futlility
    - Do not increase sample size (so lose power)
    - Maintain power under alternative (so inflate maximal sample size)
  - J = 2 optimally spaced analyses; optimal symmetric boundaries
    - Maintain power under alternative (inflate N_J, but optimize ASN)

6

6

Exploring Group Sequential Designs

- Examining operating characteristics
  - Stopping boundaries
    - Z scale
    - Conditional power under hypothesized effects
    - Conditional power under current MLE
    - Predictive power under flat prior
  - Estimates and inference
    - MLE (Bias adjusted estimates suppressed for space)
    - 95% CI properly adjusted for stopping rule
    - P value properly adjusted for stopping rule
  - Power at specified alternatives
  - Sample size distribution (as function of true effect)
    - Maximal sample size
    - Average sample size

7

7

O'Brien-Fleming Efficacy: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	--	--	--	--	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

8

8

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	--	--	--	3.955	(1.16, 5.72)	0.003
1.0	1.977	(0.00, 3.93)	0.025	1.977	(0.00, 3.93)	0.025

9

9

O'Brien-Fleming Efficacy: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.999
1.96	0.496	0.960
2.80	0.797	0.896
3.24	0.898	0.847
3.92	0.974	0.755

10

10

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	--	--	--	3.943	(1.16, 5.70)	0.003
1.01	1.977	(0.00, 3.92)	0.025	1.977	(0.00, 3.92)	0.025

11

11

O'Brien-Fleming Efficacy: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	1.005
1.96	0.499	0.966
2.80	0.799	0.901
3.24	0.900	0.851
3.92	0.975	0.758

12

12

Take Home Messages 1

- Introduction of a very conservative efficacy boundary
  - Minimal effect on power even if do not increase max N
  - Minimal increase in max N needed to maintain power
- Ease and importance of evaluating a stopping rule
- Even before we start the study, we can consider
  - Thresholds for early stopping in terms of estimated effects
  - Inference corresponding to stopping points
  - Conditional and predictive power under various hypotheses
- We can judge a stopping rule by comparing it to a fixed sample test and look at the tradeoffs between
  - Increase in maximal sample size
  - Decrease in average sample size
  - Changes in unconditional power

13

13

O'Brien-Fleming Symmetric: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.000	0.500	0.003	0.024	2.796	0.500	0.997	0.976
1.0	1.977	--	--	--	1.977	--	--	--

14

14

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.000	(-1.76, 2.80)	0.375	3.945	(1.15, 5.71)	0.003
1.0	1.973	(0.00, 3.94)	0.025	1.973	(0.00, 3.94)	0.025

15

15

O'Brien-Fleming Symmetric: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.749
1.96	0.495	0.919
2.80	0.795	0.883
3.24	0.897	0.840
3.92	0.974	0.752

16

16

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.51	0.00	(-1.75, 2.78)	0.375	3.920	(1.14, 5.67)	0.003
1.01	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

17

17

O'Brien-Fleming Symmetric: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.758
1.96	0.500	0.930
2.80	0.800	0.893
3.24	0.900	0.848
3.92	0.975	0.758

18

18

Take Home Messages 2

- Introduction of a very conservative futility boundary
- Again, minimal effects on power and/or max N
- Dramatic improvement in ASN under the null
- Conditional and predictive power thresholds are surprising
  - CP_{alt} = 0.50 for the extremely conservative OBF boundary
    - But the CI has already eliminated 3.92 with high confidence
  - CP_{est} = 0.003 and PP_{flat} = 0.024 are both very low thresholds

19

19

O'Brien-Fleming & Futility: J = 2

- Introduce two evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.5	0.331	0.644	0.017	0.068	2.776	0.500	0.997	0.975
1.0	1.963	--	--	--	1.963	--	--	--

20

20

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.5	0.468	(-1.30, 3.27)	0.228	3.925	(1.13, 5.69)	0.003
1.0	1.963	(0.00, 3.98)	0.025	1.963	(0.00, 3.98)	0.025

21

21

O'Brien-Fleming & Futility: J = 2, N = 1

- Introduce two evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.684
1.96	0.492	0.886
2.80	0.791	0.869
3.24	0.893	0.830
3.92	0.972	0.747

22

22

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.52	0.461	(-1.28, 3.22)	0.228	3.867	(1.11, 5.61)	0.003
1.03	1.934	(0.00, 3.92)	0.025	1.934	(0.00, 3.92)	0.025

23

23

O'Brien-Fleming & Futility: J = 2, Power

- Introduce two evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.705
1.96	0.504	0.914
2.80	0.803	0.892
3.24	0.901	0.850
3.92	0.975	0.762

24

24

Take Home Messages 3

- More aggressive futility boundary better addresses ethical issues associated with ineffective drugs
- I often find that sponsors are willing to accept this futility bound without increasing the sample size
- But the minimal increase in maximal sample size would seem more appropriate to me

25

25

O'Brien-Fleming & Futility: J = 4

- Introduce four evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.25	-1.108	0.719	0.000	0.008	3.976	0.500	0.999	0.999
0.50	0.321	0.648	0.015	0.063	2.811	0.500	0.997	0.977
0.75	1.258	0.592	0.142	0.177	2.295	0.500	0.907	0.874
1.00	1.988	--	--	--	1.988	--	--	--

26

26

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.25	-2.216	(-4.71, 1.74)	0.846	7.951	(4.00, 10.5)	0.000
0.50	0.454	(-1.60, 3.31)	0.263	3.976	(1.14, 6.04)	0.003
0.75	1.452	(-0.36, 3.85)	0.053	2.650	(0.30, 4.48)	0.013
1.00	1.988	(0.00, 4.06)	0.025	1.988	(0.00, 4.06)	0.025

27

27

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.580
1.96	0.478	0.783
2.80	0.776	0.761
3.24	0.882	0.723
3.92	0.966	0.650

28

28

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.27	-2.141	(-4.55, 1.68)	0.846	7.682	(3.86, 10.1)	0.000
0.54	0.439	(-1.55, 3.20)	0.263	3.841	(1.10, 5.84)	0.003
0.80	1.403	(-0.34, 3.72)	0.053	2.561	(0.29, 4.33)	0.013
1.07	1.920	(0.00, 3.92)	0.025	1.920	(0.00, 3.92)	0.025

29

29

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.622
1.96	0.504	0.840
2.80	0.803	0.808
3.24	0.902	0.762
3.92	0.975	0.680

30

30

Take Home Messages 4

- Effect of adding more analyses
  - Greater loss of power if maximal sample size not increased
  - Greater increase in maximal sample size if power maintained
  - But, improvement in average efficiency
- Can also use this example for guidance in how to judge thresholds for conditional and predictive power
  - The same threshold should not be used at all analyses
  - It is not, however, clear what threshold should be used
    - I look at tradeoffs between average efficiency and power
  - We can look at optimal (on average) designs for more guidance

31

31

Efficient: J = 2

- Introduce two optimally spaced analyses to minimize ASN
  - Type 1 error of 0.025
- Stopping boundary table

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.43	0.573	0.818	0.049	0.141	2.776	0.182	0.951	0.859
1.00	2.129	--	--	--	2.129	--	--	--

32

32

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.50	0.808	(-0.82, 3.58)	0.129	3.112	(0.34, 4.74)	0.014
1.18	1.960	(0.00, 3.92)	0.025	1.960	(0.00, 3.92)	0.025

33

33

Efficient: J = 2, Power

- Introduce two optimally spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.685
1.96	0.500	0.900
2.80	0.805	0.847
3.24	0.904	0.788
3.92	0.975	0.685

34

34

Take Home Messages 5

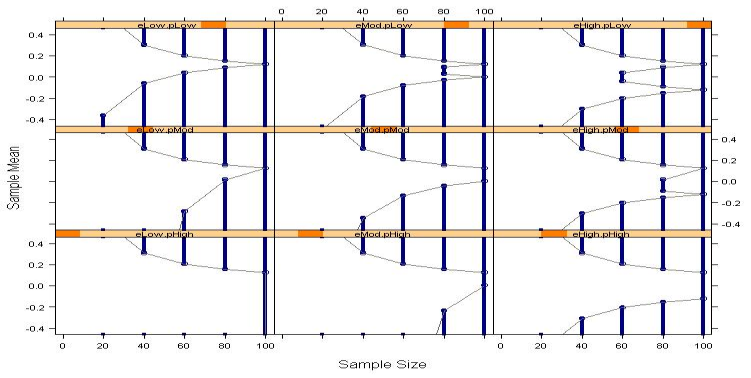
- Optimal spacing of analyses not quite equal information
- Optimal early conservatism close to a Pocock design
  - In unified family, OBF has P=1, Pocock has P= 0.5
  - Optimal P= .54
- With two analyses, increase maximal N by 18% over fixed sample
  - ASN decreases by about one third
- Again, the thresholds to use for conditional or predictive power are not at all clear
- Search for best designs should include many candidates
  - Examine many operating characteristics

35

35

Unified Family: MLE Scale

- Down columns: Early stopping vs no early stopping
- Across rows: One-sided vs two-sided decisions

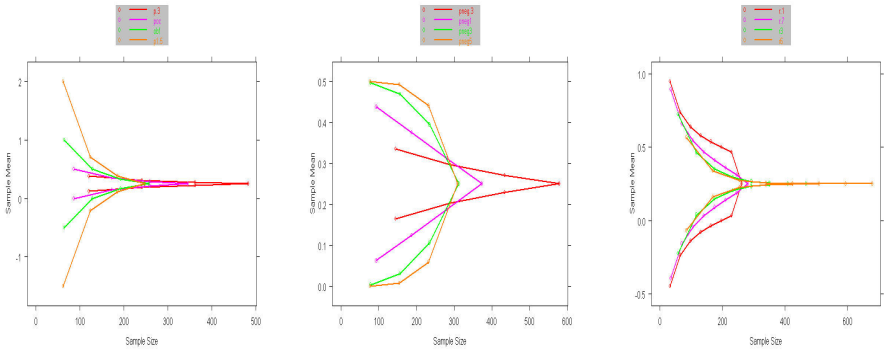


36

36

Unified Family: MLE Scale

- All of the rules depicted have the same type I error and power to detect the design alternative



37

Impact on Sampling Density

- When using a stopping rule, the sampling density depends on exact stopping rule
- This is obvious from what we have already seen.
- A fixed sample test is merely a particular stopping rule:
  - Gather all N subjects' data and then stop

38

38

Compared to Fixed Sample

- The magnitude of the effect of the stopping rule on trial design operating characteristics and statistical inference can vary substantially
- Rule of thumb:
  - The more conservative the stopping rule at interim analyses, the less impact on the operating characteristics and statistical inference when compared to fixed sample designs.

39

39

Reasons for Early Stopping

- Efficacy, Futility, Harm
- Ethical
  - Individual
    - Protect patients on study
    - Protect patients who might be accrued to study
  - Group
    - Promote rapid discovery of new treatments
- Economic
  - Avoid unnecessary costs of RCT
  - Facilitate earlier marketing

40

40

Role of Futility Boundaries

- When clinically relevant improvement has been convincingly ruled out and no further useful information to be gained
  - (Is further study of subgroups or other endpoints still in keeping with informed consent?)
- Futility boundaries usually do not indicate harm
- Because most RCT do not reject the null hypothesis, the major savings in early stopping are through a futility boundary
  - Also, not as much need for early conservatism

41

41

Potential Issue

- Compared to a stopping rule with no futility boundary the critical value at the final analysis can be lower
- Some of the trials stopped early for futility might have otherwise been type I errors at the final analysis
- Depends on the early conservatism of the futility boundary

42

42

Nonbinding Futility

- Some clinical trialists believe that FDA requires that the futility rule be ignored when making inference
  - Such builds in conservatism
  - True type I error is smaller than nominal
  - True power is smaller than normal
- This is purposely using the wrong sampling density
  - Not good statistics—game theory must be motivation

43

43

Correct Inference

- The statistically correct, efficient approach is to base inference on the real futility boundary
- Demands correct pre-specification of the futility boundary
- Demands a clear paper trail of analyses performed

44

44