Using Novel Designs in Phase II and III Clinical Trials

Elizabeth Garrett-Mayer, PhD
Division of Biostatistics
SKCCC

June 14, 2006

Outline of Talk

- Goals of Phase II study
- Single arm studies
  - Traditional (frequentist)
  - Bayesian designs
  - Multiple outcome designs
- Two (or more) arm studies
  - Traditional randomized Phase II
  - Novel multi-arm Phase II
- Goals of Phase III study
  - Characteristics
  - Complications

Goals of Phase II Trials

- Provide initial assessment of efficacy or 'clinical activity'
  - Screen out ineffective drugs
  - Identify promising new drugs for further evaluation
- Further define safety and toxicity
  - Type
  - Frequency

Important Design Considerations in Phase II trials

- Minimize cost of the trial
  - **Minimize number of patients exposed to an ineffective treatment**
  - **Enroll as few patients as "necessary" to show benefit or failure**
- Choice of patient population
  - Historical control information known?

Standard Single Arm Phase II Study

- Single arm:
- Comparison is "fixed" constant
- Binary endpoint (clinical response vs. no response)
- Often one-sided test

- Simple set-up:

$$\alpha = 0.10$$

$$\beta = 0.10 \,(\text{power} = 0.90)$$

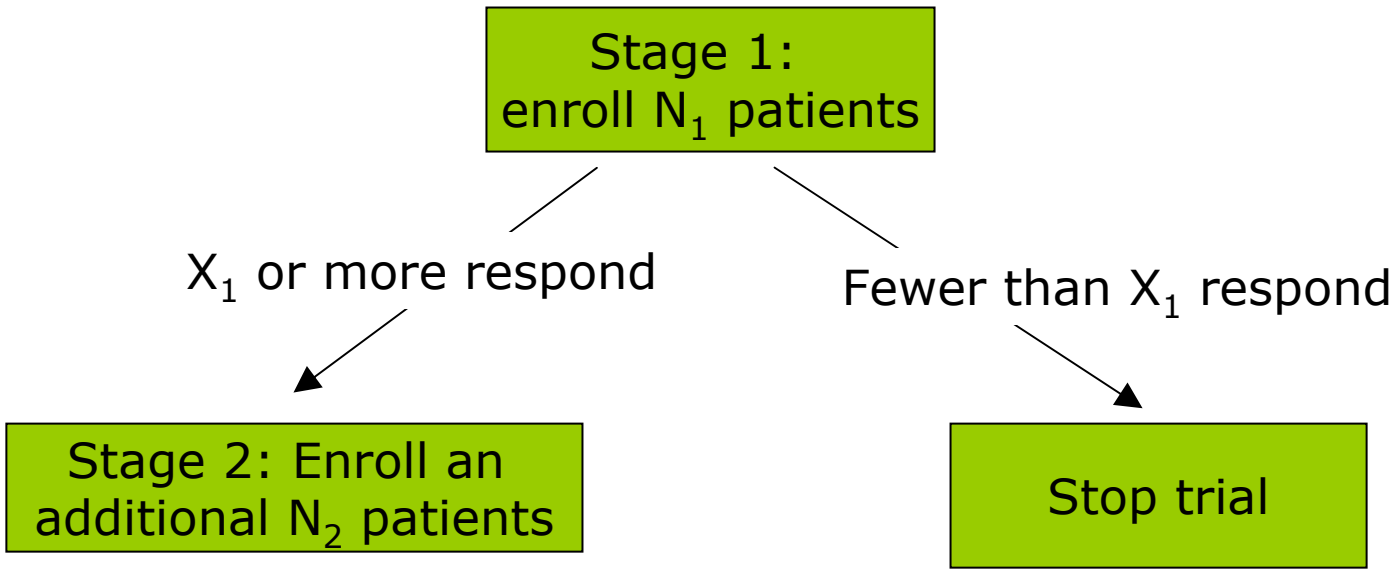
$$H_0 : p = 0.20 \,(\text{null reponse rate})$$

$$H_1 : p = 0.40 \,(\text{target response rate})$$

- Based on design parameters:
  - N=39
  - Conclude effective if 12 or more responses (i.e., observed response rate of ≥0.31)

Two-Stage Designs

- *What if by the 15th patient you've seen no responses?*
- *Is it worth proceeding?*
- Maybe you should have considered a design with an early stopping rule
- Two-stage designs:



Revised Design

- Stage 1: enroll 19 patients
  - If 4 or more respond, proceed to stage 2
  - If 3 or fewer respond, stop
- Stage 2: enroll 20 more patients (total N=39)
  - If 12 or more of total respond, conclude effective
  - If 11 or fewer of total respond, conclude ineffective
- Design properties?

$$\alpha = 0.10$$

$$\beta = 0.10 \,(\text{power} = 0.90)$$

$$H_0 : p = 0.20 \,(\text{null reponse rate})$$

$$H_1 : p = 0.40 \,(\text{target response rate})$$

Multi-Stage Designs

- Two-stage
  - Simon two-stage (1989)
    - Used in example
    - **MANY "optimal" designs**
    - Preserves alpha and power, and permits early look
  - Gehan two-stage (1961)
    - At stage 1, stop if 0 responses
    - Choose N_1 such that early stopping has 'good' properties
    - "Special case" of Simon two-stage

Multi-Stage Designs

- Three-stage
  - Ensign et al. (1994)
    - Permits early stopping when a moderately long sequence of initial failures occurs.
    - Two opportunities for early stopping
    - Our example: stop trial if $\leq$2/18 responses or $\leq$ 9/33
    - Not used as commonly as two-stage

Early Stopping

- FUTILITY stopping
- The designs discussed so far ONLY allow stopping if there is strong evidence that the treatment is not efficacious
- Can also have early stopping for efficacy
  - Generally not popular
  - Important to accumulate evidence to support claim of efficacy
  - But, not stopping prolongs time to launch phase III

Frequentist versus Bayesians

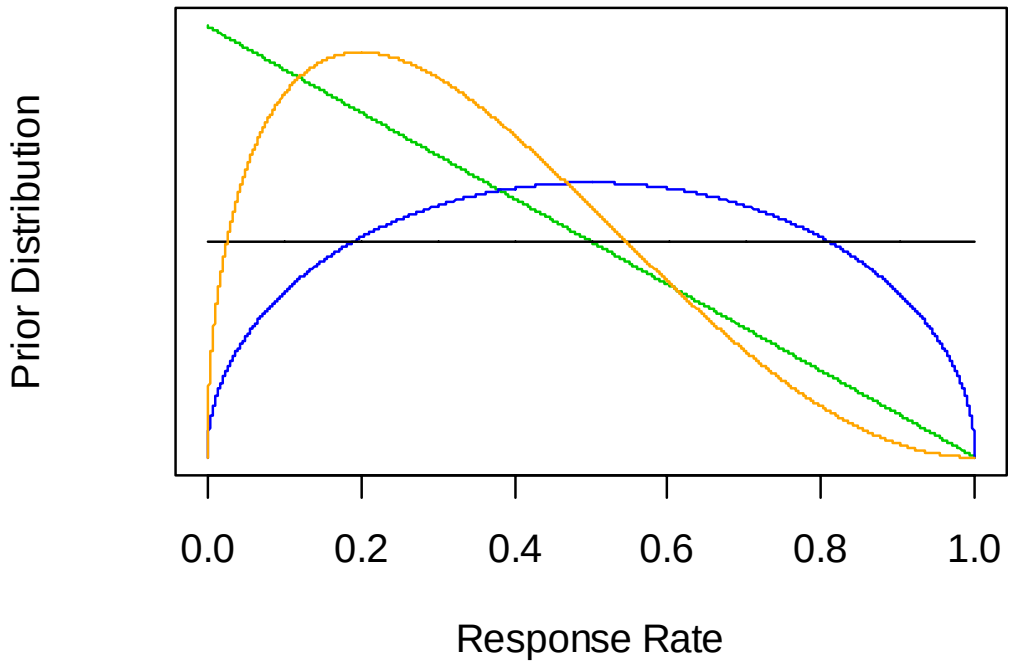
- So far, "frequentist" approaches
- Frequentists: α and β errors
- Bayesians:
  - Quantify designs with other properties
  - General philosophy
    - Start with prior information ("prior distribution")
    - Observe data ("likelihood function")
    - Combine prior and data to get "posterior" distribution
    - Make inferences based on posterior

Bayesian inference

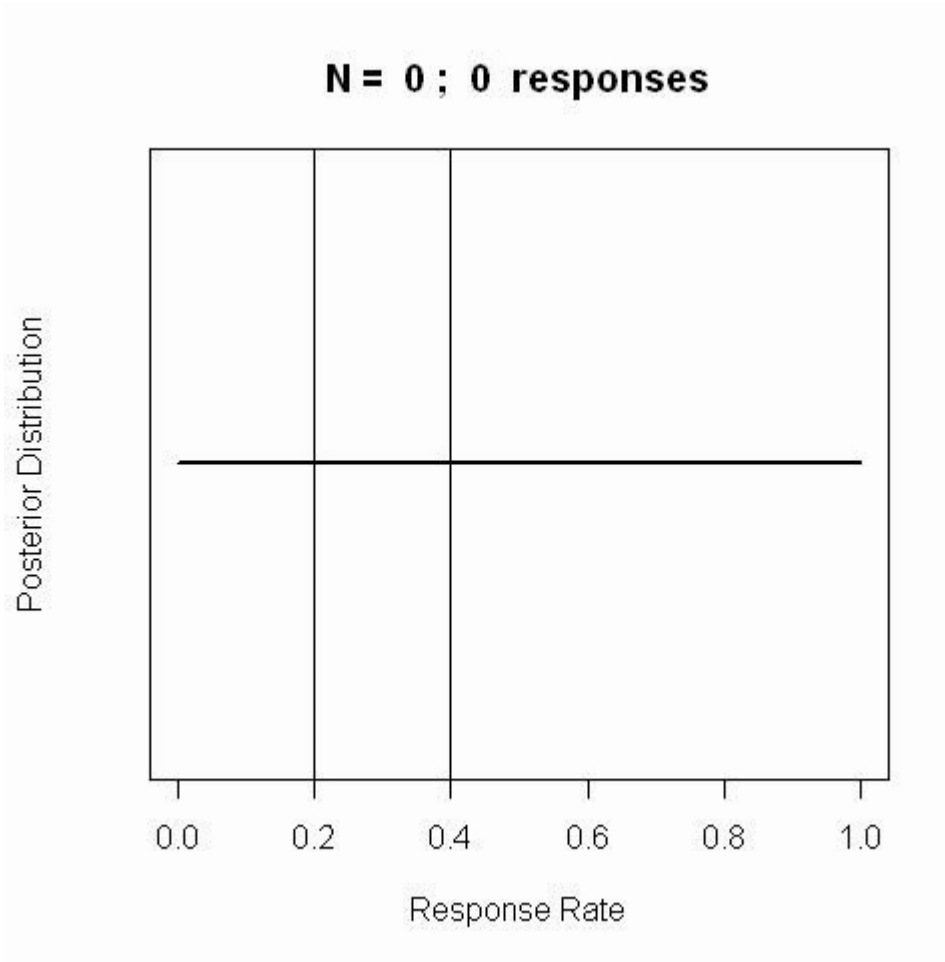
- No p-values and confidence intervals
- From the posterior distribution:
  - Posterior probabilities
  - Prediction intervals
  - Credible intervals
- Bayesian designs
  - Can look at data as often as you like (!)
  - Use information as it accumulates
  - Make “what if?” calculations
  - Helps decide to stop now or not

Bayesian Designs

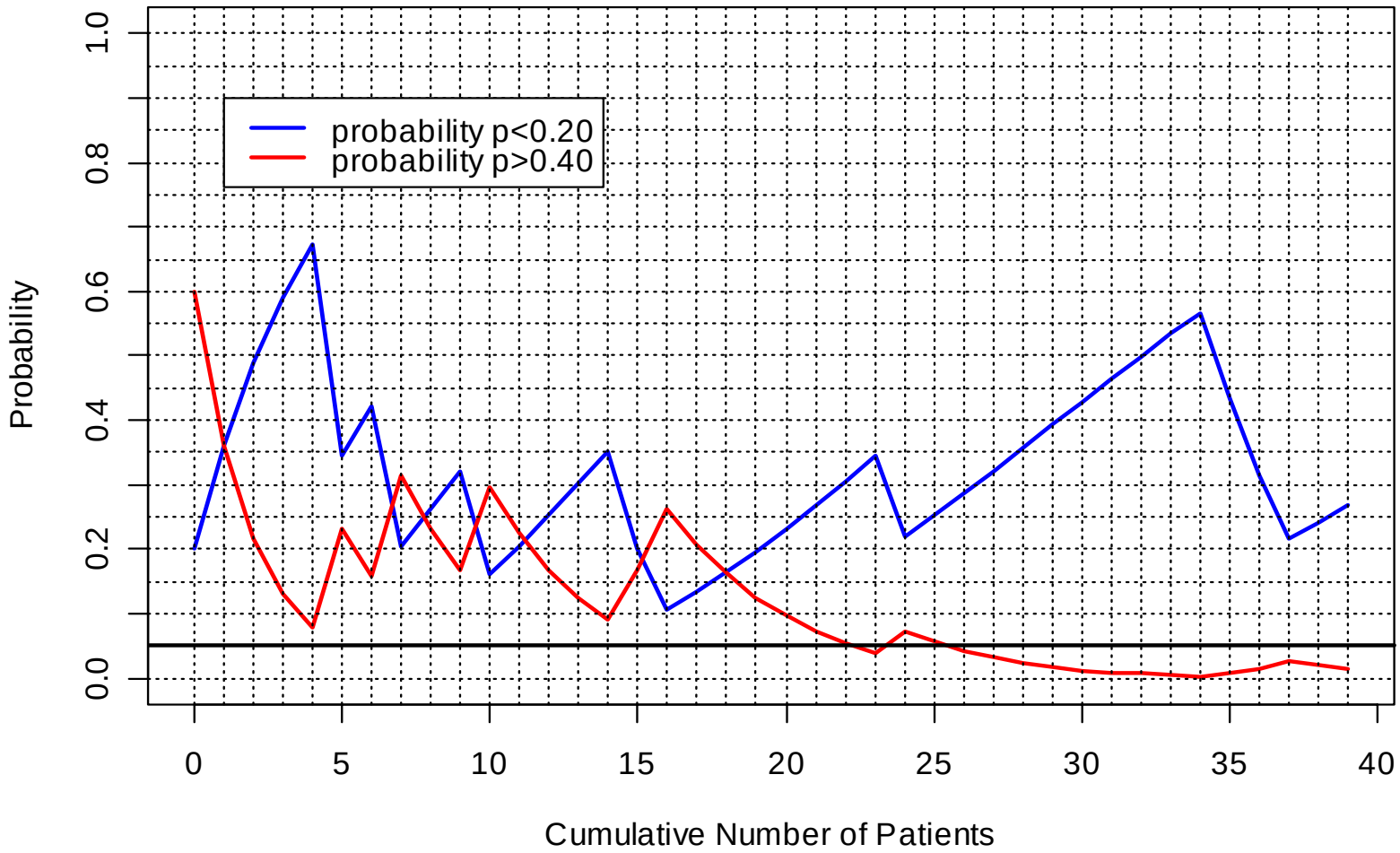
- Requires 'prior'
  - Reflects uncertainty about the response rate
  - Can be 'vague', 'uninformative'
  - Can be controversial: inference may change



Bayesian design example

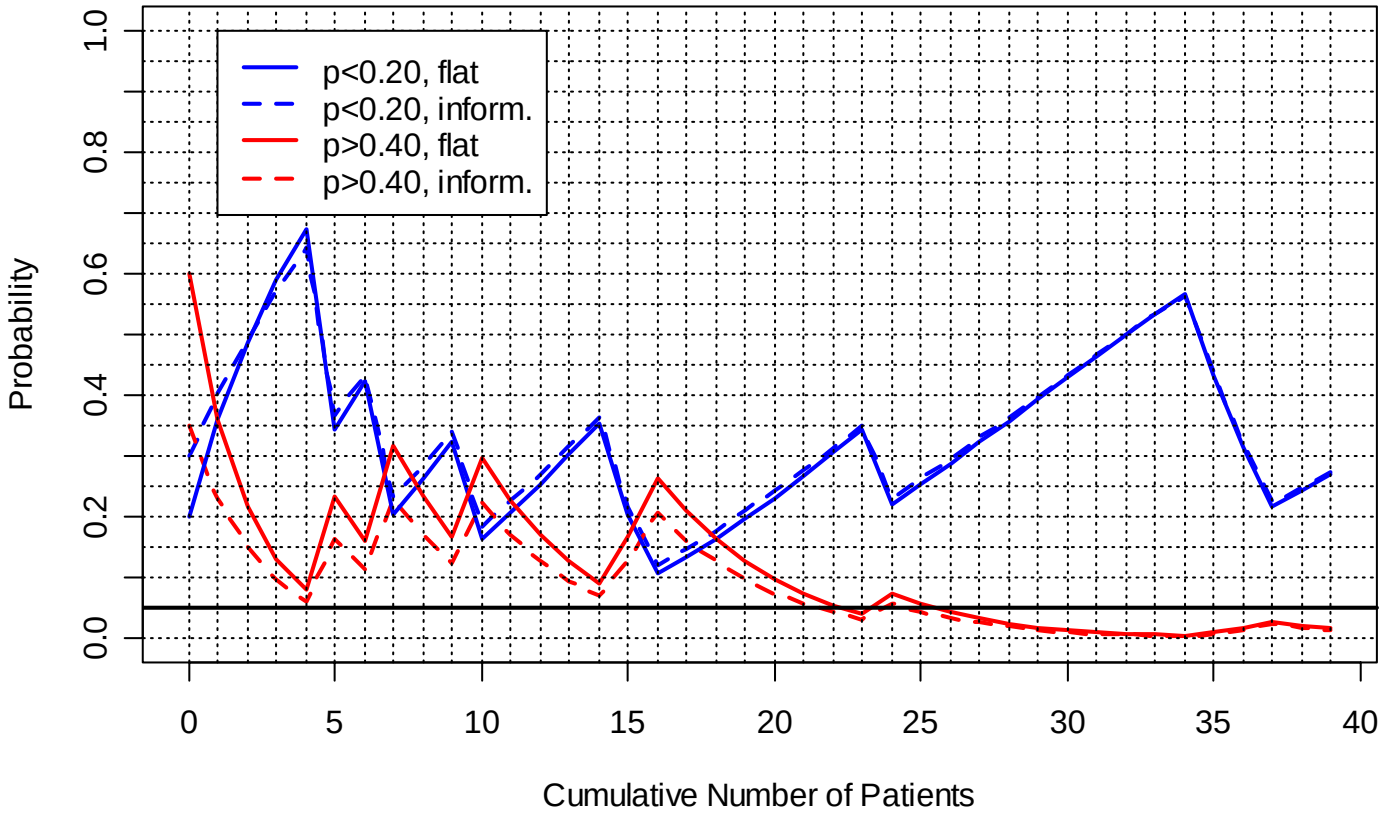


Posterior Probabilities



Other priors

- What if we had used a different prior?
- Assume informative "orange" prior



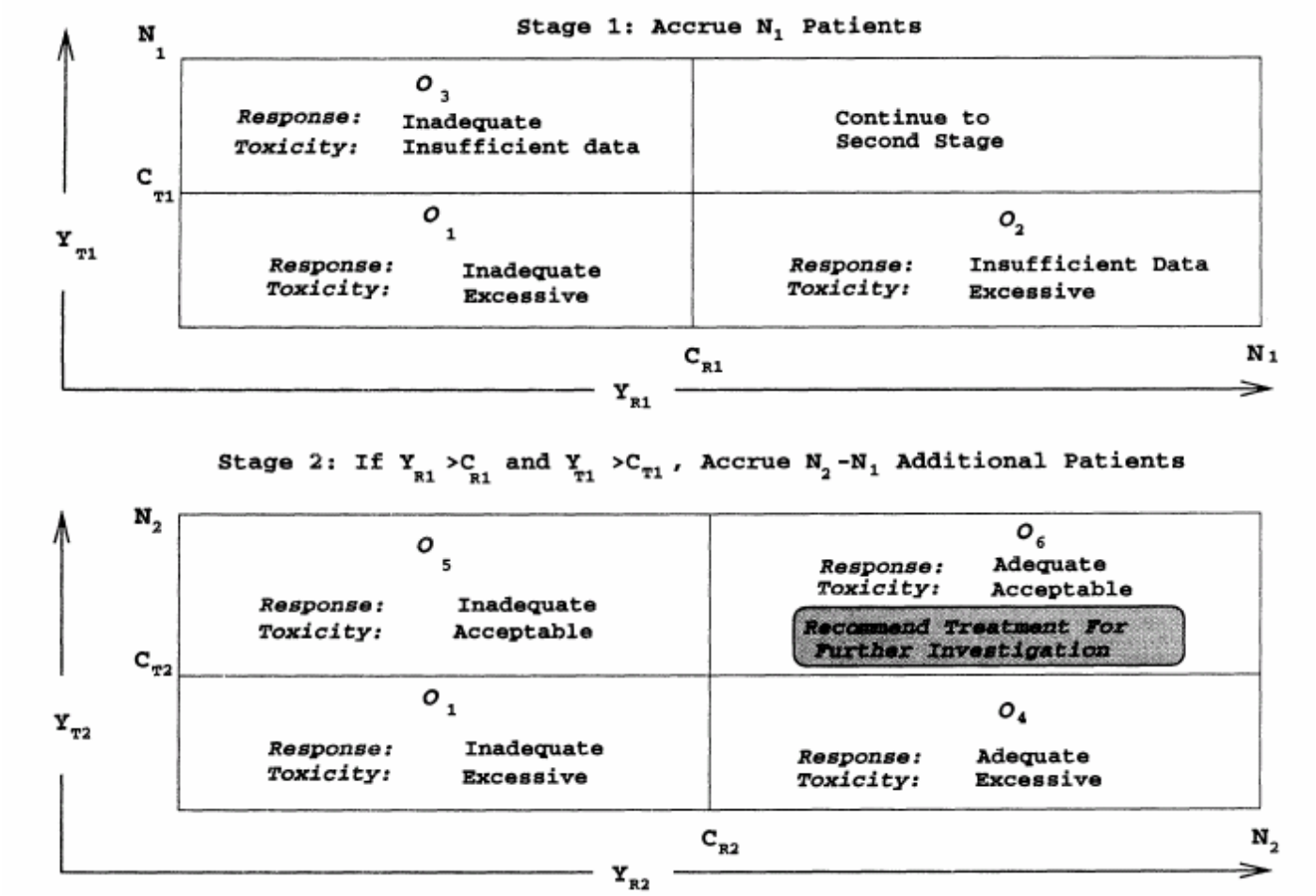
Likelihood Approach

- Similar to Bayesian
- Royall (1997), Blume (2002)
- No prior distribution required
- Quantified by intuitive properties
  - instead of α and β
  - "Probability of misleading evidence"
  - (i.e. choosing the wrong hypothesis)
- Likelihood ratio used for making inferences
- Can look at data as it accumulates

Multiple Outcomes

- Phase II = "safety + efficacy" trial
- Then why are we only talking about efficacy?
- Bryant and Day (1995): extend Simon two-stage to incorporate both outcomes
- Thall and Cheng (1999): treated as "true" bivariate outcome

Bryant and Day Design

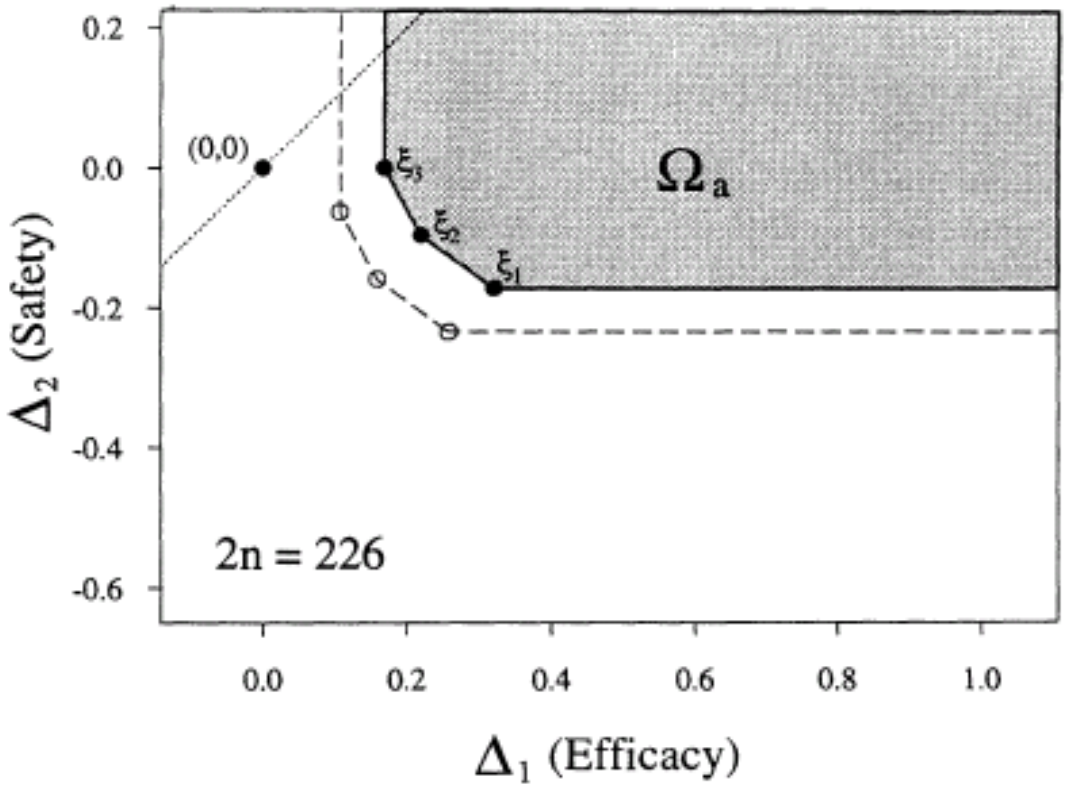


Examples of Bryant and Day Designs:

Criterion				Optimal Designs					
P_{R0}	P_{R1}	P_{T0}	P_{T1}	N_1	C_{R1}	C_{T1}	N_2	C_{R2}	C_{T2}
0.05	0.25	0.60	0.80	22	1	14	43	4	29
0.10	0.30	0.60	0.80	21	2	13	46	7	31
0.20	0.40	0.60	0.80	24	5	15	54	14	36
0.30	0.50	0.60	0.80	23	7	14	57	21	28
0.40	0.60	0.60	0.80	25	10	15	53	25	36
0.60	0.80	0.60	0.80	20	12	12	49	33	33

- Example (first row)
  - null rates: efficacy 5% and safety 60%
  - alternative rates: efficacy 25% and safety 80%
  - Stage 1: enroll 22 patients
    - stop if (1) one or fewer responses OR (2) 14 or fewer "safe" patients
  - Stage 2: enroll an additional 21 patients (total N=43)
    - conclude a negative study if (1) four or fewer responses OR (2) 29 or fewer "safe" patients

Thall and Cheng Design



Other "novel" issues

- Time to event outcomes in Phase II
  - Response rate no longer the 'outcome of choice' in Phase II studies
    - targeted agents may not shrink cancer
    - we're learning: tumor shrinkage ≠increased survival
  - Time to event outcomes more common
    - time to progression
    - time to relapse
    - time to death
  - More than ever, need early stopping
  - Simon's two-stage does not apply
  - Bayesian and Likelihood methods are becoming more appealing

Summary of Single Arm Phase II Trials

- STRONGLY CONSIDER ALLOWING FOR EARLY STOPPING
- Bayesian and likelihood designs:
  - Allow early stopping as soon as strong evidence develops
  - More complicated to implement
    - High-maintenance: many analyses
    - Computationally intensive
  - For Bayesian: choice of prior can be tricky
  - Lack of objectivity and potential loss of "equipoise"
- Frequentist designs:
  - Usually just one interim analysis
  - Simple implementation

Why randomized phase II?

- Classic phase II studies:
  - Single arm study where results are compared to historical control rate.
  - Problem: this is not always 'satisfying'
    - Requires patient populations to be comparable
    - Might not have information to derive control rate (e.g. disease progression is of interest and not response rate)

- Comparative randomized studies (phase III):
  - Allow us to compare two arms
  - Problem:
    - Large sample size (more than twice a single arm study)
    - Costly
    - Large undertaking based on scant preliminary data

Why randomized phase II?

- Want to explore efficacy
- Not willing to invest in phase III (yet)
- Want some "control" or "prioritization"
- Primarily two different kinds of randomized phase II studies
  - *Phase II selection design (prioritization)*
  - *Phase II designs with reference control arm (control)*
- Also phase II/III studies

Common design of randomized phase II study

- Two parallel one arm studies (classic case)
- **Do not directly compare arms to each other.**
- Compare each to "null rate"
- Example: null response = 0.20, alternative response=0.40, alpha=0.10 (one-sided), power=0.90.
  - Two parallel one-arm studies:
    - *Test each treatment to see if it is better than null rate*
    - *For two arm study, need N=78 patients (39 per arm)*
  - Comparative study:
    - *Test to see if one treatment is better than the other treatment*
    - *For two arm study, need N=160 patients (80 per arm)*

Classic Randomized Phase II designs

- Phase II selection designs (Simon, 1985)
  - "pick the winner"
  - 90% chance of choosing better arm so long as true difference in response rates is >15%.
  - Appropriate to use when:
    - Selecting among NEW agents
    - Selecting among different schedules or doses
  - NOT appropriate when
    - Trying to directly compare treatment efficacies (not powered)

Classic Randomized Phase II designs

- Phase II selection designs (continued)
  - Uses 2+ Simon two-stage designs
    - Each arm is compared to a null rate
    - Must satisfy efficacy criteria of Simon design
    - Move the "winner" to phase III
    - Only have to pick winner if more than one arm shows efficacy
  - Can be used when the goal is prioritizing which (if any) experimental regimen should move to phase III when no a priori information to favor one.

Classic Randomized Phase II designs

- Randomized Phase II designs with reference arm
  - Includes reference arm to ensure that historical rate is "on target"
  - Reference arm is not directly compared to experimental arm(s) (due to small N)
  - Can see if failure (or success) is due to incomparability of patient populations
  - Problem: if it turns out that historical control rate used is very different from what is observed in reference arm, then trial should be repeated (Herson and Carter, 1986)

Phase II/III studies

- Several versions {Schaid (1988), Storer (1990), Ellenberg and Eisenberger (1985), Scher and Heller (2002)}
- General idea
  - Begin with randomized phase II study
  - Randomize to control arm & experimental arm(s)
  - If some threshold of efficacy is met, continue to phase III sample size for direct comparison
- Benefits:
  - Allow use of phase II data in phase III inference
  - Minimize delay in starting up phase III study
  - Uses concurrent control
- Cons:
  - The sample size for the phase II part is approximately twice as large as would be needed for standard phase II
  - Need phase III infrastructure developed even if it stops early.
- **Would be useful if MOST phase II studies showed efficacy**
- Really, these could be considered phase III designs with very aggressive early stopping rules.

Other Randomized Phase II designs?

Lots of randomized studies are calling themselves randomized phase II studies these days:

- If outcome of interest is surrogate
  - Correlative (biomarker)
  - Clinical (response)
- If sample size is relatively small but direct comparison is made
- If study is comparative, but is not definitive for whatever reason (e.g. if α and β are large, patient population can not be generalized)

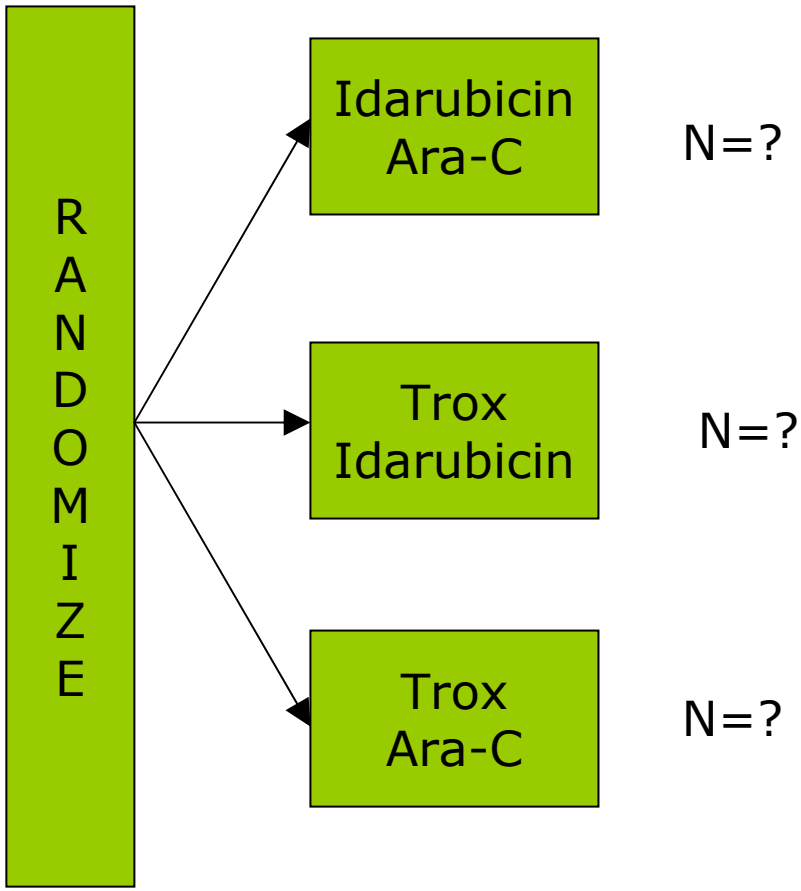
Adaptive Randomization Designs

- Randomization is "adapted" based on accumulated information
- Risk factors: (Halpern and Brown, 1986)
  - Want balance in two treatment arms with respect to stage of disease, age, gender, etc.
  - Instead of random allocation, assign biasing towards "balance"
  - Works well when there are **many** stratification factors
  - Both Bayesian and frequentist approaches
  - Common example: (Bayesian) Biased coin

Adaptive Randomization Designs

- (continued)
- Outcome (Bayesian/Likelihood)
  - Assign treatments according to accumulated information about best treatment. (Berry and Eick, 1995)
  - Assign with higher probabilities to better therapies
  - Example: Troxacitabine in AML (Giles et al. 2003)

Adaptive Designs



Adapt the randomization to learn while effectively treating patients on trial:

(1)Begin by randomizing with equal chance per arm
(2) Then, adjust probability of assignment to reflect the knowledge of the best treatment

Adaptive Designs

- Summary of trial results:
  - TI dropped after 24th patient
  - Trial stopped after 34 patients (TA dropped)

Complete responses by 50 days

IA	10/18 = 56%
TA	3/11 = 27%
TI	0/5 = 0%

Summary of Multi-arm Phase II trials

- Think about why/whether a multi-arm trial is needed
- Very useful when there is lack of historical data for comparison
- Phase II randomized is NOT a short-cut to avoid a larger more definitive trial
- Adaptive designs can be very efficient for selection, but require more maintenance

Goals of Phase III studies

- Compare one or more treatments to get definitive efficacy evidence
- Obtain sufficient evidence to convince the FDA to approve an experimental treatment
- Determine if treatment showing preliminary efficacy “works” on gold standard outcome (e.g. survival)

Phase III trials: Comparative

- Almost ALWAYS multi-arm
- Almost ALWAYS multi-center in oncology
  - Co-operative groups
  - Pharma studies
- Sample size in the range of 100's to 1000's
- BIG undertaking
  - That is why we need to be VERY EFFICIENT AND SMART in our Phase II evaluations

Phase III trials

- DSMC
- Bayesian approach: Adaptive randomization
  - Pros: early stopping, efficient treatment, may be more attractive to patients
  - Cons: high maintenance, many analyses, statistician dependent, not considered 'prime-time' designs
- Frequentist approach: Interim analyses
  - 'alpha-spending': similar to phase II, but not as elegant
  - Not as efficient: can only stop at predefined timepoints although strong evidence may have accumulated earlier

Other complications of Phase III trials

- Cross-over
- Differential drop-out
- Differential compliance (not such an issue in most oncology trials)
- Survival is gold-standard outcome
  - What about accounting for treatment received ***after*** progression?
- "ITT":
  - what does this mean?
  - How does it apply (differently) in oncology trials?
- Time-to-event outcomes: interval-censored
  - Results depend on intervals at which you look

Summary: Issues with innovative designs

- Statistically intensive
  - "buy your statistician a beer (or bourbon)"
  - Probably cannot be used "off-the-shelf"
  - require specialized software
- Need to be validated
  - do they behave as promised?
  - are they 'robust' (i.e., do they work when incorrect assumptions are made)?

References (1)

- Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. Stat Med. 1995 Feb 15;14(3):231-46.
- Blume, JD. Likelihood Methods for Measuring Statistical Evidence, Stat Med. 2002 (21), 2563-2599.
- Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. Biometrics. 1995 Dec;51(4):1372-83.
- Ellenberg SS, Eisenberger MA. An efficient design for phase III studies of combination chemotherapies. Cancer Treat Rep. 1985 Oct;69(10):1147-54.
- Ensign LG, Gehan EA, Kamen DS, Thall PF. , An optimal three-stage design for phase II clinical trials.
  Stat Med. 1994 Sep 15;13(17):1727-36.
- Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. J Chronic Dis. 1961 Apr;13:346-53.
- Giles FJ, Kantarjian HM, Cortes JE, Garcia-Manero G, Verstovsek S, Faderl S, Thomas DA, Ferrajoli A, O'Brien S, Wathen JK, Xiao LC, Berry DA, Estey EH. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. J Clin Oncol. 2003 May 1;21(9):1722-7.

References (2)

- Halpern J, Brown BW Jr. Sequential treatment allocation procedures in clinical trials--with particular attention to the analysis of results for the biased coin design. Stat Med. 1986 May-Jun;5(3):211-29.
- Herson J, Carter SK. Calibrated phase II clinical trials in oncology. Stat Med. 1986 Sep-Oct;5(5):441-7.
- Royall R. *Statistical Evidence:A Likelihood Paradigm,* London, Chapman & Hall, 1997.
- Schaid DJ, Ingle JN, Wieand S, Ahmann DL.A design for phase II testing of anticancer agents within a phase III clinical trial. Control Clin Trials. 1988 Jun;9(2):107-18.
- Scher HI, Heller G. Picking the winners in a sea of plenty. Clin Cancer Res. 2002 Feb;8(2):400-4.
- Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treat Rep. 1985 Dec;69(12):1375-81.
- Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials. 1989 Mar;10(1):1-10.
- Storer BE. A sequential phase II/III trial for binary outcomes. Stat Med. 1990 Mar;9(3):229-35.
- Thall PF, Cheng SC. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. Biometrics. 1999 Sep;55(3):746-53.