DF/HCC
DANA-FARBER / HARVARD CANCER CENTER

Considerations for Choosing Among Types of Phase II Designs

Paul Catalano

July 23, 2010

Purpose of Phase II Studies

- Determine if a new agent or a new treatment regimen appears sufficiently efficacious to be worth further investigation
  - Not definitive
  - Not powered for all outcome endpoints
- Verify the safety of the therapy
- Provide statistical rigor/formal evaluation context and targeted patient population
  - More formal than expansion cohort of a phase I

Phase II Designs

- Single arm with single analysis point
- Single arm with interim stopping rules (usually with suspension of accrual)
- Randomized parallel arms design
- Randomized selection designs ('pick-the-winner')
- Comparative randomized control
- Randomized discontinuation designs

Important Definitions

- Type I error rate / α : False positive rate
  Chance inactive regimen moves forward
- Type II error rate / β : False negative rate
  Chance active regimen does not move forward
- Power = 1 – β

General Phase II Approach

- Often formulate as testing a statistical null hypothesis (H_0) versus an alternative (H_a)
  - e.g. H_0: $p_r = 0.05$ vs H_a: $p_r = 0.20$

    where p_r is the true proportion of patients who will respond to the new agent
- Consequence of a type I error (α): an ineffective agent will be studied further
  - Use $\alpha = 0.10$ (one-sided)
  - Often larger than in phase III studies

General Approach

- Consequence of a type II error (β): an effective agent will not be studied further
  - β should be ≤ 0.10
- In practice, tend to be multiple phase II studies performed in multiple diseases and centers, so the overall chance of missing an effective treatment is lower
- Selection of therapies for phase III testing is based on available data, not usually on a single phase II study

Classic Design for Evaluating New Agents

- Patients refractory to standard therapy
- If some patients improve (respond), agent must have some activity
- Often consider H_0: $p_r = 0.05$ vs. H_a: $p_r = 0.20$
- Simon's (1989) optimal two-stage designs minimize expected sample size under H_0

Classic Design for Evaluating New Agents

- Simon's optimal design for $p_r = 0.05$ vs 0.20
  - 1st stage: treat 12 patients; stop if no responses
  - 2nd stage: treat 25 patients; conclude inactive if < 4 / 37 (11%) respond
- Optimal designs like these attractive for new agents in diseases without prior evidence of activity
  - 4/37 = 11% is enough evidence to reject H_0

Single Stage Accrual Designs

- Might be appropriate
  - If some prior evidence of activity
  - For combinations of new drugs with standard treatments
- Example: H_0: $p_r = 0.20$ vs H_a: $p_r = 0.37$ (null rate depends on level of activity for standard treatment)
  - Single stage: 45 patients, reject H_0 if 13 / 45 (29%) or more respond

Two-stage designs

H_0: $p_r = 0.20$ vs H_a: $p_r = 0.37$

1st stage: accrue 25 patients,
Inactive if <5/25 respond

2nd stage: accrue 25 more patients
Inactive if 13/50 or fewer respond

True H_0	Assumed H_0	H_a	Power	Type I error
0.20	0.20	0.37	0.92	0.11
0.25	0.20	0.37	0.92	0.35
0.15	0.20	0.37	0.92	0.01

With fixed cut points for success in each stage

Two-stage designs

- Consider power fixed at 92% and type I error at 11%
- Sample sizes needed for optimal designs:

H_0	H_a	Approximate Sample Size
0.20	0.37	50
0.25	0.37	95
0.15	0.37	40

Improvement in Disease Stabilization

- Cytostatic agents and targeted therapies might improve disease stabilization rates rather than improve response rates
- Test for improvement in disease stabilization rates
  e.g. H_0: $p_s = 0.30$ vs. H_a: $p_s = 0.50$, where p_s = proportion stable or responding (free of progression) at x months (e.g. x = 4)
- Calculations the same as for response, in principle (there are catches, though)

Other Endpoints

- TTP or PFS (not at a specific time point)
  - Kaplan-Meier estimate at single time or other nonparametric estimation
  - Parametric (e.g. exponential) models can be slightly more efficient (tricky though to anchor hypotheses)
- Multinomial: For example, test
  H_0: $p_r = 0.05$ and $p_s = 0.30$ vs.
  H_a: $p_r > 0.05$ or $p_s > 0.30$
  - Less efficient than single endpoint (binomial)
  - May be more difficult to interpret
- Survival generally not appropriate

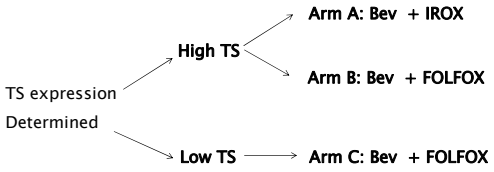
Types of Randomized Phase II Designs

- Separate evaluation of each arm (parallel)
  - Each arm evaluated in a similar population
- Selection designs: select the 'best' arm for further study
- Comparative randomized control
- Randomized discontinuation

Randomized designs are larger and more complex – need to explain each arm to patients

Randomized parallel arms design

E4203: Frontline Metastatic CRC



E4203

- Primary endpoint RR (CR+PR) on arms A, B, C
- High TS Expression
  - Test Ho: RR $\leq$ 28% vs. Ha: RR $\geq$ 45%
  - Two-stage design with 69 eligible patients *per arm*
  - Stage 1: 12 or more responses out of 38 accrued
  - Stage 2: 26 or more total responses among 69 pts
  - 90% power under alternative, type I error 0.05
- Low TS Expression
  - Test Ho: RR $\leq$ 54% vs. Ha: RR $\geq$ 70%
  - Stage 1: 21 or more responses out of 36 accrued
  - Stage 2: 55 or more total responses among 87 pts
  - 90% power under alternative, type I error 0.05

Control Arms in Phase II Setting

- Concern about selection bias in studies without a simultaneous control group
  - Studies can enroll different patient groups even with the same nominal population.
  - Population drift, stage migration, changes in supportive care are all factors
  - Historical rates hard to determine (but critical w/o randomization!)
- Control groups more appropriate for evaluating contribution to a combination or effect on progression than for determining if *any* response activity
- Comparing studies from different groups is very difficult; setting single arm nulls also very hard
- Situation especially difficult with PFS as the endpoint

Control Arms (cont'd)

- Often not needed because
  - Phase II studies can only detect fairly large effects, so biases would need to be large
  - Consequence of a false positive is further testing of an inactive drug (would find out, but only after potentially high cost)
  - Where relevant, cooperative group or other studies conducted in the same network with central data review produce fairly consistent results
- Increase the time, expense and resources for phase II evaluation (especially for 'screening')
- Single arm trials are operationally easier

Issues to consider when entertaining control arms

- Definition of endpoints can vary across trials (especially for endpoints like PFS)
- Supportive care and other factors can change and impact null hypotheses
- Entry criteria for studies impact assessment of null
- Increase sample size
- Need to carefully consider alternative
- Could still end up with imbalances across arms
- Studies take longer, harder to consent

Impact on sample size

- H_0: $p_r = 0.20$ vs. H_a: $p_r = 0.37$
- Two-stage single arm design required 50 patients
- Power of 92%
- 1-sided type I error rate of 11%
- A randomized design with a comparative control would require 215 patients.
- If H_0 is really 0.25, then 215 patients provides only 70% power
- To limit sample size to ~100 patients, the alternative would need to be $p_r = 0.45$.

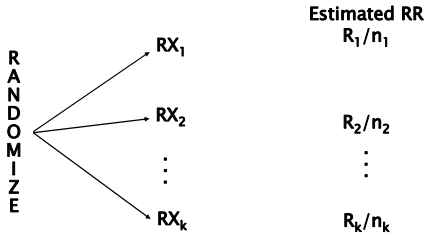
Randomized Selection (Pick-the-Winner) Designs

- Randomize between 2 or more experimental arms (no control arm)
  - In a sense, least efficacious arm is a control for the others
- Select the best arm for further evaluation
- Usually define 'best' to be the arm with the numerically best observed outcome, no matter how small the difference
- Careful statistical design needed, including efficacy rules for each arm

Randomized Selection (Pick-the-Winner) Designs

- With two arms, $\alpha \cong 0.50$
  - Rationale: doesn't matter which arm is selected if they are nearly equivalent
- Should incorporate separate efficacy tests for each arm, too
  - 1-stage or 2-stage
- Usually prefer randomizing over a series of separate studies
  - Facilitates (informal) comparisons
  - Guards against sampling bias

Randomized Selection (Pick-the-Winner) Designs



RX_j is 'winner' if $R_j/n_j > R_i/n_i$ for $i \neq j$

Can use other endpoints

Randomized Selection (Pick-the-Winner) Design

- Example: 2-stage design for H_0: $p_r = 0.20$ vs H_a: $p_r = 0.40$ enrolls 17 patients in the 1st stage and 20 in the 2nd ($\alpha = \beta = .10$)
- Apply this design to each arm in a 2-arm randomized selection design

True RR		Prob arm is winner		
p_{r1}	p_{r2}	RX1	RX2	Neither
.20	.40	.015	.890	.095
.30	.40	.147	.758	.095

Randomized Selection (Pick-the-Winner) Designs

- Probability of selecting the best arm declines as the number of arms increases
- $X_1 \sim Bin(50, .32)$; $X_2 \ldots X_k \sim Bin(50, .20)$ gives $P\{X_1 > max(X_2,\ldots,X_k)\} = .90$ for $k = 2$ and $P\{X_1 > max(X_2,\ldots,X_k)\} = .72$ for $k = 6$
- Advanced renal trial of several targeted agents: 6 arms, n=55 / arm
  - TTP compared via Cox model
  - If one arm has median TTP of 7.2 months and the other 5 have median TTP of 4.8 months (50% improvement), then the probability of selecting the best arm is 0.87

What's not appropriate for pick-the-winner designs?

- Trials of the form A versus A+B
  - Why should control arm A be considered as a winner against a new agent?
  - These are not randomized comparative designs
  - Need to be comfortable at outset moving forward with any of the arms
- Agents with carry-over effect
  - Dilute treatment effect of randomization

Comparative Randomized Control Design

- Discussed for evaluating cytostatic agents in Korn et al. (2001)
- Randomize experimental vs. standard and formally compare the arms
- Appropriate if don't have a reasonable prior estimate of expected control arm outcomes
- Endpoint could be any of the standard phase II endpoints (e.g. TTP, response)
- Might target larger differences than a phase III

Comparative Randomized Control Design

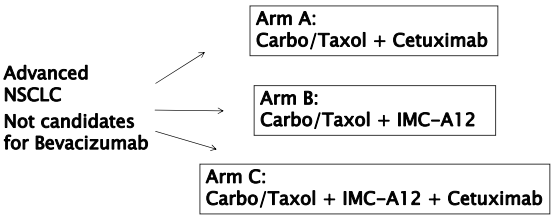
- Test provides a suggestive (phase II) evaluation with a larger α (e.g. 0.10 to 0.20)
  - Appropriate for screening new agents
  - If positive, still needs to be followed by a definitive phase III study
    - Large type I error; no power for survival
  - Korn et al. suggest using $\alpha = 0.20$, because the sample size with $\alpha = 0.10$ is large enough that it might be better to go directly to the definitive study

The cost of randomization

- Return to our example of 0.20 vs 0.37 (recall single arm design required 50 patients)
- One-sided type I error rate of 0.10
- 90% power with log rank test
- Target a PFS HR of 0.62

PFS rates	Medians (mos)	Sample Size	Follow-up
6 mos	2.6 vs. 4.2	120	12 mos
12 mos	5.2 vs. 8.4	140	12 mos
24 mos	10.3 vs. 16.7	160	18 mos

Another Example: E4508



E4508

- 180 patients over 12 months with 10 additional months of follow-up
- 88% power for A vs C, B vs C
- 1-sided 0.10 significance level
- Increase in median PFS from 3.5 months to 5.6 months
- HR = 0.625
- No plan to compare the single agent arms A and B

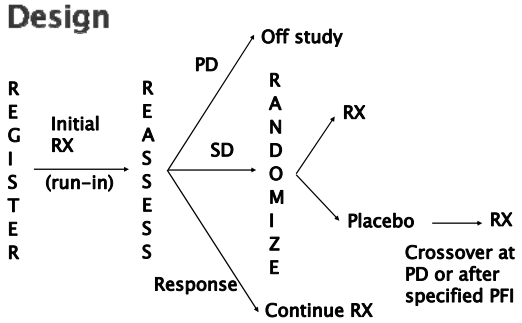
E4508 – sensitivity to the null

- If null median PFS rate is 4.0 mos not 3.5 mos:
  - Power decreases to 67%
  - Targeting HR of 0.71
- If null median PFS rate is 3.0 mos not 3.5 mos:
  - Power increases to 97%
  - Targeting HR of 0.54

Randomized Discontinuation Design

- An enrichment strategy based on randomizing patients who appear to be doing well on the treatment
- Initially all patients are treated, patients free of progression for some period of time are randomized between continuing treatment and placebo, with crossover from placebo to treatment at progression or specified PFI
- Complex design with a blinded randomization and 3 registration points

Randomized Discontinuation Design



Randomized Discontinuation Design

- Usefulness depends on success of run-in in selecting patients benefiting from treatment
  - TTP is highly variable in most diseases, so randomized population will be a mixture
  - Korn et al. (2001), Capra (2004) suggest often less efficient than standard RCT
- Carry-over effect could dilute difference between randomized arms
- Requires much larger sample size

Main Points

- In many settings, conventional phase II designs may still be appropriate
- Start-up costs, accuracy of assumptions for single-arm two-stage designs are a concern
- Randomized phase II studies allow evaluation of multiple agents or schedules and protect against sampling bias
- They also can provide internal control to protect null
- Selection designs are useful for informal comparison and identifying promising combinations

Main Points

- Control arms can be effective in some settings, but can't ignore costs
- Control arms don't give a free ride on all design assumptions
- Survival is seldom (never?) the best phase II endpoint and sample size prohibitive
- Randomized discontinuation designs may not be appropriate and need to be strongly justified

DF/HCC

References

- Capra WB (2004). Comparing the power of the discontinuation design to that of the classic randomized design on time-to-event endpoints. Controlled Clinical Trials 25:168-177.
- Freidlin B, Dancey J, Korn EL, Zee B, Eisenhauer E (2002) Multinomial phase II trial designs (letter to the editor). Journal of Clinical Oncology 20:599.
- Fu P, Dowlati A, Schluchter M. Comparison of Power Between Randomized Discontinuation Design and Upfront Randomization Design on Progression-Free Survival . JCO Sep 1 2009: 4135-4141.
- Gan HK, Grothey A, Pond GR, Moore MJ, Siu LL, Sargent D. Randomized phase II trials: inevitable or inadvisable? J Clin Oncol. 2010 May 20;28(15):2641-7.

DF/HCC

References

- Gray R, Manola J, Saxman S, Wright J, Dutcher J, Atkins M, Carducci M, See W, Sweeney C, Liu G, Stein M, Dreicer R, Wilding G, DiPaola RS. Phase II clinical trial design: methods in translational research from the Genitourinary Committee at the Eastern Cooperative Oncology Group. Clin Cancer Res. 2006 Apr 1;12(7 Pt 1):1966-9.
- Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC (2001). Clinical trial designs for cytostatic agents: are new approaches needed? Journal of Clinical Oncology 19:265-272.
- Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. J Clin Oncol. 2005 Jul 1;23(19):4450-7.
- Rosner GL, Stadler W, Ratain MJ (2002). Randomized discontinuation design: application to cytostatic antineoplastic agents. Journal of Clinical Oncology 20:4478-4484.

DF/HCC

References

- Rubinstein L, Crowley J, Ivy P, Leblanc M, Sargent D. Randomized phase II designs. Clin Cancer Res. 2009 Mar 15;15(6):1883-90.
- Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol. 2005 Oct 1;23(28):7199-206.
- Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol. 2005 Mar 20;23(9):2020-7.
- Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, Ratain MJ, Le Blanc M, Stewart D, Crowley J, Groshen S, Humphrey JS, West P, Berry D. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clin Cancer Res. 2010 Mar 15;16(6):1764-9.

DF/HCC

References

- Simon R (1989). Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials 10:1-10.
- Simon R, Wittes RE, Ellenberg SS (1985). Randomized phase II clinical trials. Cancer Treatment Reports 12:1375-1381.
- Tang H, Foster NR, Grothey A, Ansell SM, Goldberg RM, Sargent DJ. Comparison of error rates in single-arm versus randomized phase II cancer clinical trials. J Clin Oncol. 2010 Apr 10;28(11):1936-41.
- Taylor JM, Braun TM, Li Z. Comparing an experimental agent to a standard agent: relative merits of a one-arm or randomized two-arm Phase II design. Clin Trials. 2006;3(4):335-48.
- Wieand HS. Randomized phase II trials: what does randomization gain? J Clin Oncol. 2005 Mar 20;23(9):1794-5.

DF/HCC