Optimizing Phase 1 and 2 Clinical Trials Design

Stephanie Green and Tao Wang

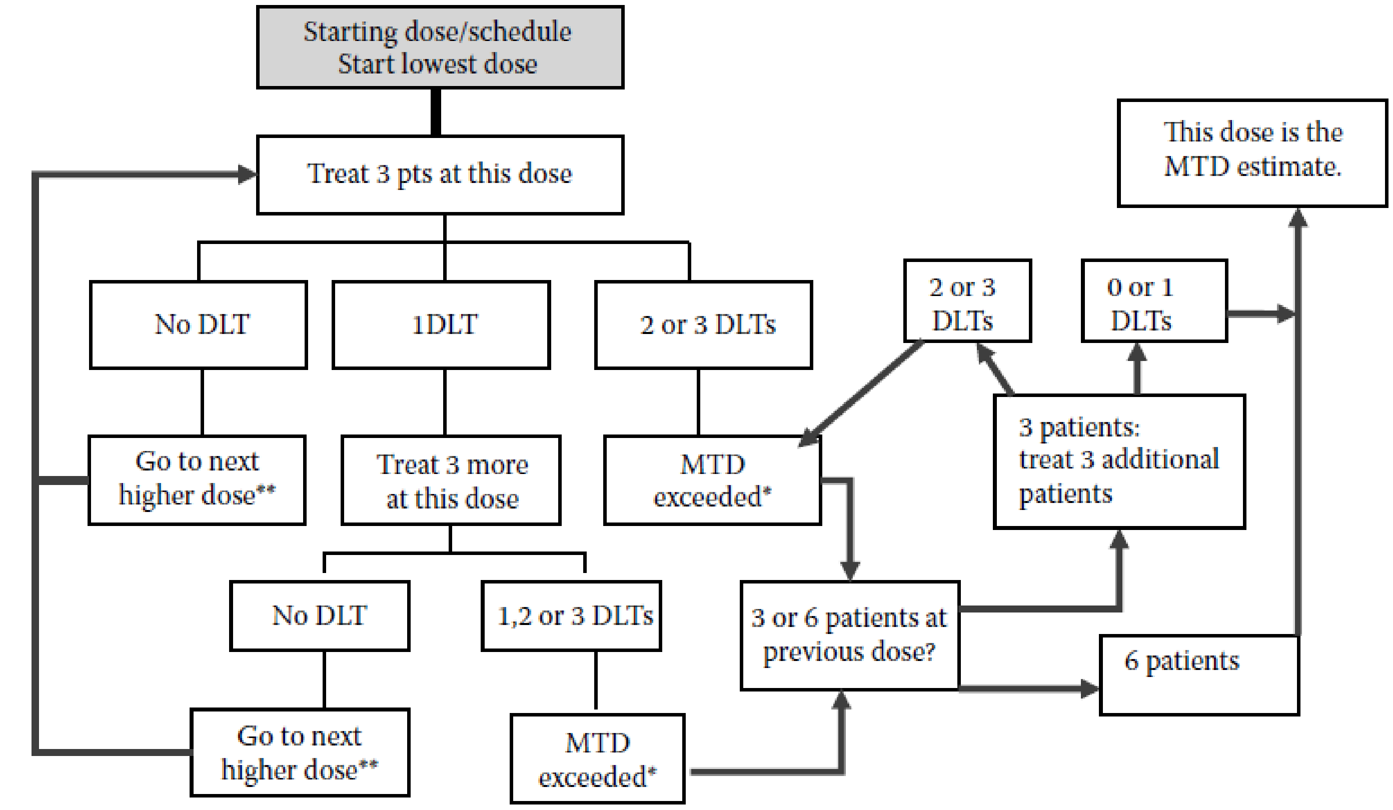
OUTLINE

- Phase 1 options
  - Improve efficiency over 3+3
  - Use other designs if standard assumptions are not met
  - Phase 1→ 2 strategy
- Phase 2 options
  - Single arm, randomized
  - Risk benefit design
  - Phase 2 → 3 strategy

Phase 1

Primary aim: identify dose for Ph III testing based on cycle 1 Dose Limiting Toxicities

3+3 design



Starting dose/schedule
Start lowest dose

Treat 3 pts at this dose

No DLT

1DLT

2 or 3 DLTs

Go to next higher dose**

Treat 3 more at this dose

MTD exceeded*

No DLT

1,2 or 3 DLTs

Go to next higher dose**

MTD exceeded*

3 or 6 patients at previous dose?

3 patients: treat 3 additional patients

6 patients

2 or 3 DLTs

0 or 1 DLTs

This dose is the MTD estimate.

*If the MTD is exceeded at the lowest dose level stop. The MTD cannot be estimated.
** If there are no higher dose levels, stop. The MTD cannot be estimated.

Classic 3+3 design

- -Inefficient: too many patients treated at low doses
- -Poor estimate of MTD: biased and sensitive to both starting dose and shape of dose-toxicity curve.
- -Inflexible: no standard modification for the design if 16%-33% isn't target toxicity level

Accelerated titration designs

Simon et al 1997

Rapid dose increase in single patients until toxicity observed, then switch to standard.

Potential Benefits

- Fewer dose steps
- Fewer patients
- Increased probability of entering a dose cohort with potential clinical benefit

One of few innovations to make inroads

Accelerated titration designs

Simulation averages for 20 trial scenarios	3+3	Dose doubling, one pt per cohort Switch to 3+3 after DLT or 2 grade 2 in Cycle 1
Sample size	39.9	20.7
# trials with average>35	10	0
# trials with average>50	6	0
Pts with worst grade 0-1	23.3	3.9
Pts with worst grade 3-4	7.4	11.1

Up and down designs

Storer, 1989:

- Escalate single pts until DLT observed (or other single pt scheme), then accrue in 3s.

  -increase if no DLTs

  -same dose if 1 pt with DLT

  -decrease if 2-3 with DLTs
- Stop and estimate after fixed sample size using a logistic model: $f(d) = \exp(\alpha+\beta d)/(1+ \exp(\alpha+ \beta d))$

Up and down designs

Estimate of MTD is improved

Proportion treated at low doses reduced

Small increase in proportion treated at unacceptably high doses.

Further improvement if accumulated information is used to inform next dose (Ivanova et al 2003)

such as K in a row designs:

- Decrease if DLT
- Increase if no DLTs in the last K patients
- Otherwise treat at same dose

Estimate at end of trial using isotonic regression

Model based designs

Model based, as opposed to algorithmically based, designs require a precise definition of true MTD.

The dose-toxicity relationship is a function

f(d)=population probability of DLT at dose d.

The MTD is the dose *d** for which f(*d**)=p_t, where p_t is the target probability of DLT

Continual reassessment

Single parameter model f(d,λ) O’Quigley et al 1990)

Estimate of MTD updated after each patient

Next pt treated at dose closest to new MTD estimate

Multiple ways to implement:

-One patient at a time vs. 2-3.

Generally small efficiency loss, more stable

-Single stage vs. two stage.

Increase rapidly until toxicity noted, then switch to CRM

Continual reassessment

Fixed N vs. stopping rule, such as

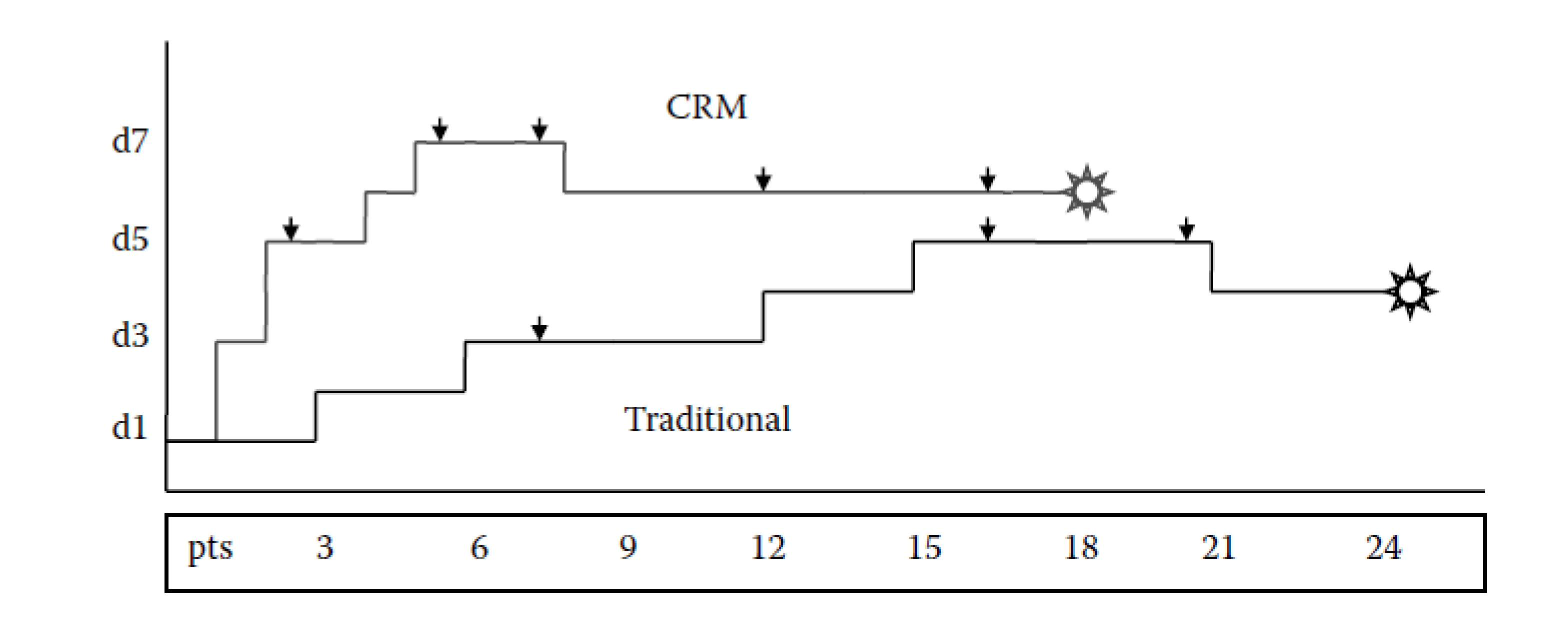
Stop if MTD settled prior to max sample size: N subjects are treated at same dose.

Use closest dose to MTD estimate vs. use lowest untested dose below MTD estimate.

Using more conservative approaches some efficiency is lost but safety is improved

Designs are often more efficient than 3+3

Continual reassessment



Hoped for benefit.

Comparisons

Target probability .33

- 3+3: Estimate targets .2, not .33. High % of patients under-dosed, very low % overdosed
- Up and down and CRM: estimates reasonably on target, modest decrease as true probability of DLT at starting dose increases. Similar low % of patients underdosed. Precision of estimate not high though better than 3+3
- CRM: Highest % overdosed, particularly if starting dose has low probability of DLT

(Storer, 2012)

Escalation with Overdose Control (EWOC)

A drawback to CRM is that f(d,λ) is not correct for any dose but d* resulting in overdose risk

EWOC (Tighiouart et al, 2005) is related to CRM, but considers probability of DLT for all doses instead of just MTD.

Overdoses are reduced compared to CRM at the expense of longer trial time

Time to Event CRM Variation

TITECRM – time to event CRM

For use when DLTs do not occur quickly, eg, RT studies. Probability of event during time period of length T is assessed.

Decisions on dose escalation/de-escalation are based on accumulating time to event information, not just on patients with complete information at time T.

Timelines are shortened and estimation accuracy is not compromised (Cheung and Chapelle, 2000).

Toxicity probability interval design

Specify intervals for probability of DLT that represent under-dosing, acceptable dosing, overdosing
If DLT estimate at current dose indicates under-dosing, increase dose in next cohort, if acceptable stay at same dose, otherwise reduce dose &/or declare unacceptable (Ji et al 2010 and modification Yuan et al, 2014)

Scheme is specified using Bayesian calculations, but does not require updates during the trial - operationally straightforward.

Toxicity probability interval design

Number treated at current dose

Number of DLTs at current dose

	1	2	3	4	5	6	…
0	↑	↑	↑	↑	↑	↑	
1	↓	=	=	=	=	↑	
2		too toxic	↓	=	=	=	
3			too toxic	too toxic	↓	=	
4				too toxic	too toxic	too toxic	
…							

Designs for other assumptions

Examples:

Underlying assumption for Phase I designs is that higher doses are both more toxic and are better with respect to efficacy, so the MTD will be the most efficacious. If this is incorrect, the previous designs are not suitable.

Combination treatments: Partial ordering expected.

Shallow dose toxicity relationship

Some biologic agents have low anticipated toxicity over a wide range of doses with responses expected to plateau instead of continuing to increase with dose. Dose escalation based on response may be reasonable in this setting (Hunsberger et al, 2009)

- Consider slope of line describing dose response relationship for 4 highest doses
- Stop escalation when slope is estimated ≤0 (response no longer increasing)

Joint assessment of toxicity and efficacy

Thall and Cook, 2004

Aim is to identify doses with acceptable toxicity and efficacy.

Define toxicity-efficacy contours: sets of probabilities of equal acceptability.

(pe, 0) = acceptable efficacy if no tox

(1, pt) = acceptable tox if 100% efficacy

(p1, p2) = a point of same acceptability

Joint assessment of toxicity and efficacy

-Treat the first cohort at a starting dose identified clinically

-After assessment of each patient use Bayesian methods to identify acceptable doses given the current data

A dose d is acceptable if calculations indicate it is reasonably likely that

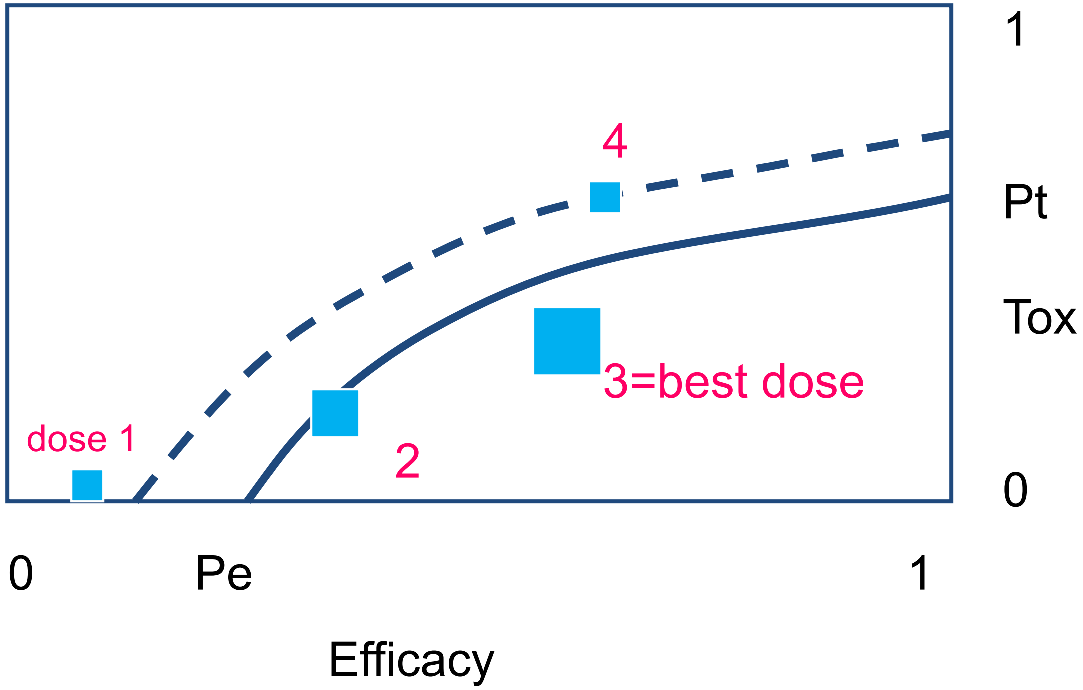
1) efficacy probability is > pe and

2) toxicity probability < pt

Joint assessment of toxicity and efficacy

- If there are acceptable doses, the next cohort is treated either at the next untried dose level or at the most desirable level (dose on the most desirable contour) whichever is lower.
- If there are no acceptable doses then trial is closed and no dose is selected.
- If the trial is not stopped early and there are acceptable doses remaining at the end of the trial then the most desirable of the remaining doses is chosen for further study.

Joint assessment of efficacy and toxicity



Joint assessment of toxicity and efficacy

Simulations are promising

-Superior to previous proposals in this setting

-Modest sample size when no doses are acceptable

-Rapid identification of a dose in the setting of increasing then decreasing efficacy as a function of dose.

Biomarker version also proposed (Bekele and Shen, Biometrics 2005

Two agent designs

Up and down for two agents. (Ivanova & Wang, 2004)

Goal is to find a set of MTD combinations.

For target probability of MTD=.3, assign pts in groups of 2. Next dose is a function of %DLTs at the current doses (p) and of DLT in the most recent cohort:

No DLT and p<.3: ↑agent 1, =agent 2

No DLT and p>.3: ↓agent 1, ↑agent 2

DLT and p>.3: ↓agent 1, =agent 2

DLT and P<.3: =agent 1, ↓agent 2

Two agent design

Escalate to a fixed sample size.

Outcome consistes of the estimates of the probability of DLT at each dose combination used

p11	p21		pM1
..			..
p1N	p2N		pMN

Probabilites assumed to be non-decreasing as a function of dose in both directions.

MTD combinations are identified using isotonic regression estimation to produce smoothed estimates conforming to the partial ordering.

Phase 1→ Phase 2

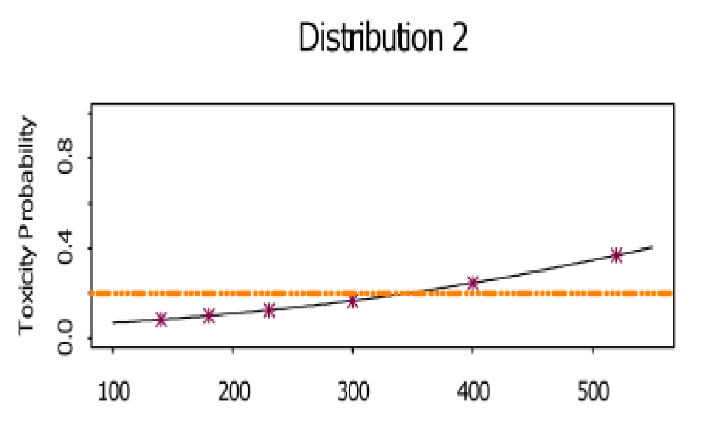
Risk of sub-optimal dose is high, with attendant consequences

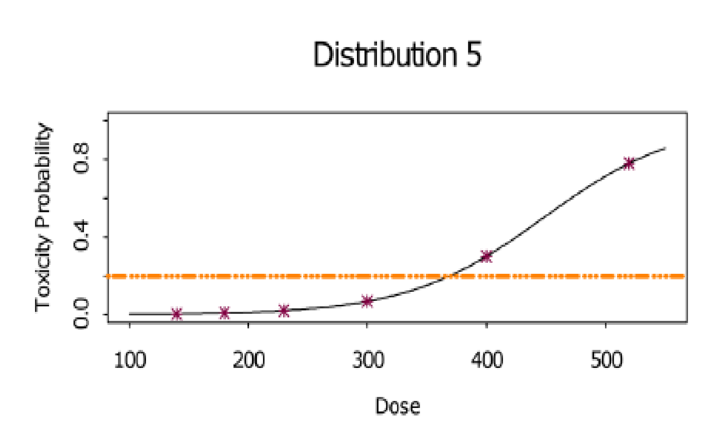
Example: Pegilated liposomal doxorubicin in ovarian cancer. MTD identified as 50, positive phase III with this dose

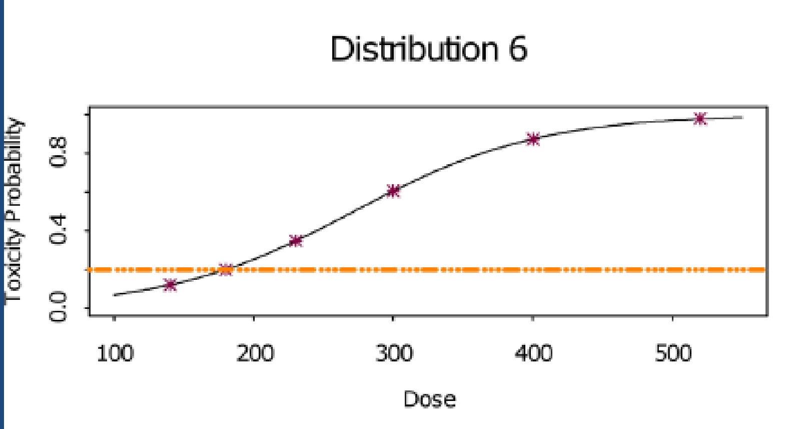
- Severe hand-foot syndrome in 25% of pts
- Dose of 40 used in practice with reduced tox, indirect evidence of similar evidence
- Regulatory requires study designs with approved dose.

Phase 1→ Phase 2

Dose selection cannot be highly accurate given small sample sizes







Comparisons

Onar-Thomas and Xiong, Contemporary Clinical Trials, 2010

Shape of dose resp curve	flat 300, 400	steep late 300, 400	steep early 180, 140
Design characteristics	CRM	CRM	CRM
Sample size	15	16	9
% with DLT	18%	14%	33%
% treated ≥2 above mtd	5%	5%	8%
Trial duration (18 pts/yr)	482	523	273
Dose chosen:			
% correct dose	30%	37%	28%
% next closest	27%	60%	41%
% disagreement with 3+3	58%	44%	36%

Phase 1→ Phase 2

Use of expansion cohorts common, typically about 10 patients at the Phase 1 MTD for additional safety information and preliminary efficacy information. However, no formal plans are specified; utility is uncertain.

Recommendation:

Implement decision criteria to improve probability of choosing the correct dose

Consider larger expansion cohorts

Phase 1→ Phase 2

Reassess MTD at end of expansion at MTD, or allow dose changes during expansion and reassess at end of expansion (Iasonos & O'Quigley, 2013)

Incorporate efficacy

- Allow dose changes; success = response without DLT (Ivanova,2003)
- Randomize to MTD and next lower dose, choose based on modified selection design (Hoering et al 2012)

Phase 1 Conclusions

- Old standard 3+3 targets probability of DLT ~20% regardless of intended target
- Methods using more information than results from previous cohort generally have better properties
- Alternative designs needed if assumptions are not met, eg, some biologic agents, 2-agent escalation schemes, long term DLTs
- Small sample size precludes accurate identification; make better use of expansion cohorts before final choice of dose.

Phase 2

Standard design aim—decide whether Phase III should be considered or not.

– Demonstration of antitumor activity

– Informal assessment of benefit/risk

– Sufficient evidence to anticipate reasonable likelihood of Phase III succes

Single arm design

Standard design aim—decide whether Phase III should be considered or not.

The usual set up is

H_0: p= p_0 vs H_1: p= p_A

where p_0 is the historical probability of response and p_A is a probability of response for which good power is required.

Single arm design

Various ways to choose sample size and stopping rule

- E.g., Minimizing expected sample size for given level level and power. (Simon, 1987)
  - Note this approach requires that the precise number of patients be enrolled at each stage. Otherwise, optimality is lost.
- A practical approach is to test the alternative ($p=p_A$ vs $p<p_A$) at the interim analysis at a conservative level, and the null at the .05 level at the final analysis. (Green and Dahlberg, 1992)

Single arm design

Issues

– Variability in patient population despite similar eligibility criteria; choice of null may be inappropriate

– Historical information often not well characterized

But - Sample size for randomized trial with same level and power is 4X the size of single arm

Single arm stratified design

Probability of response will vary within patients who satisfy eligibility criteria. If there are good historical estimates for common subsets a stratified design can be considered

CML Example: Three cohorts: second line, third line, previously treated advanced

Frequency ~ 40%, 40%, 20%

Historical response probabilities p1, p2, p3 ~ .2, .1, .1 with advanced group anticipated to recieve less benefit from new treatment. Alternatives = .4, .3, .2

Single arm stratified designs

Choose approximate sample size based on standard two stage two arm design using assumed cohort distribution and nulls and alternatives for each

CML example: Null = .4x.2 + .4x.1 + .2x.1 = .14

Alternative = .4x.4 +.4x.3 + .2x.2 = .32

Sample size: 40 (20 per stage)

Interim stopping only for insufficient response, stopping bound = m11xp1 + m12xp2 + m13xp3 = a1, where m1j is the attained first stage sample size for cohort j.

Second stage rejection rule is based on observed cohort sizes:

For the attained cohort sizes, adjust null and use as the test boundary the number of responses=a such that level is maintained. (Jung et al. 2012)

Single arm stratified designs

Design characteristics determined exactly:

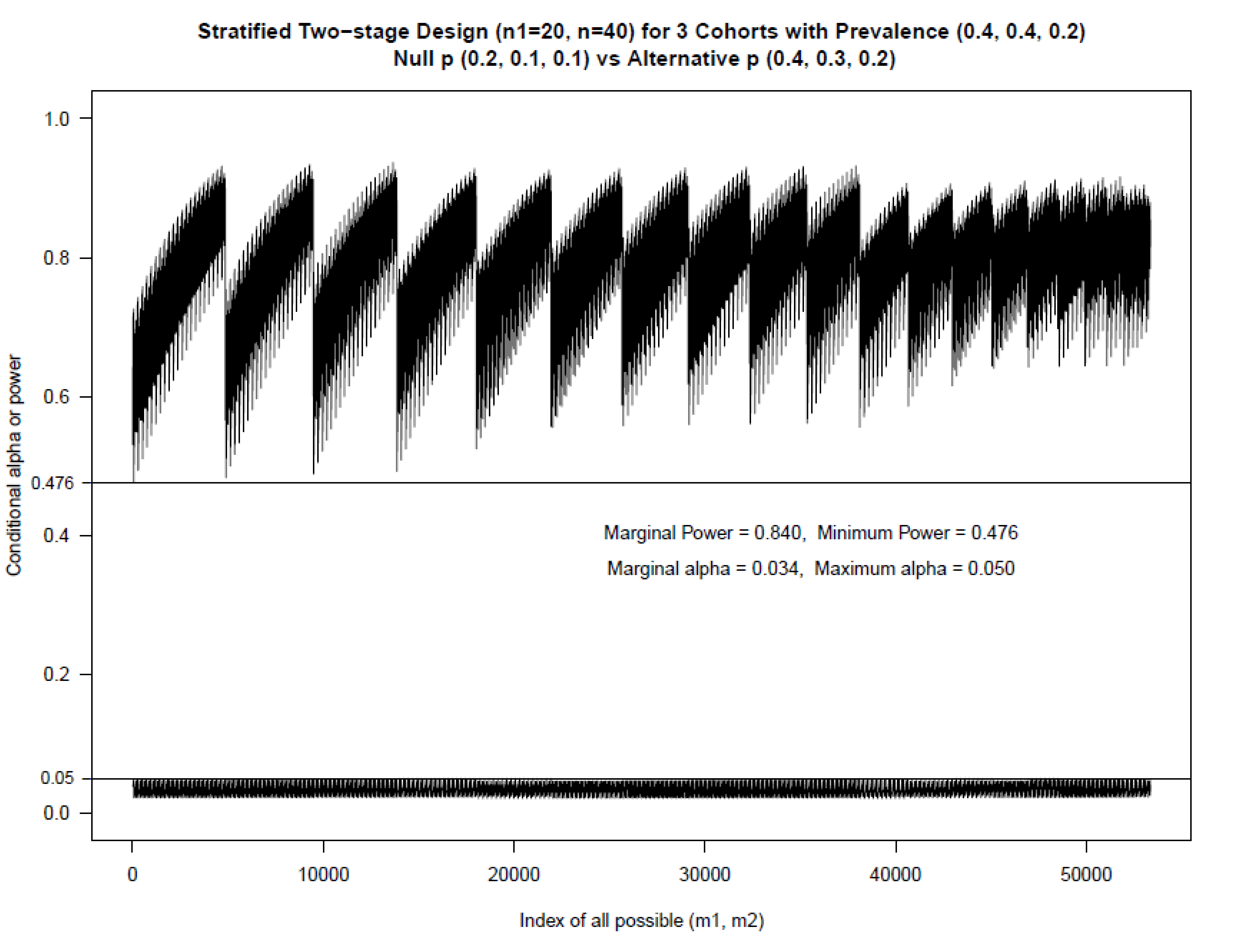
- For each 53361 possible (m1, m2, m3) determine a1 and a.
- Go to stage 2 if stage 1 responses x1 > a1
- Reject H0 if total responses x=x1+x2 > a
- Weight level and power for each by probability of (m1,m2, m3) given the assumed cohort distribution

- Design has satisfactory marginal type I error control as well as adequate marginal power
  - Marginal alpha=0.034, and Margin power=0.840

Single arm stratified design

- Within the set of (m1, m2, m3) most likely to be observed by various thresholds, the minimum power is reasonably high
  - For example, there is 97.5% probability that observed (m1, m2, m3) will have Pr(m1, m2, M3) >= 0.00001. In this set, the minimum power is 0.731.

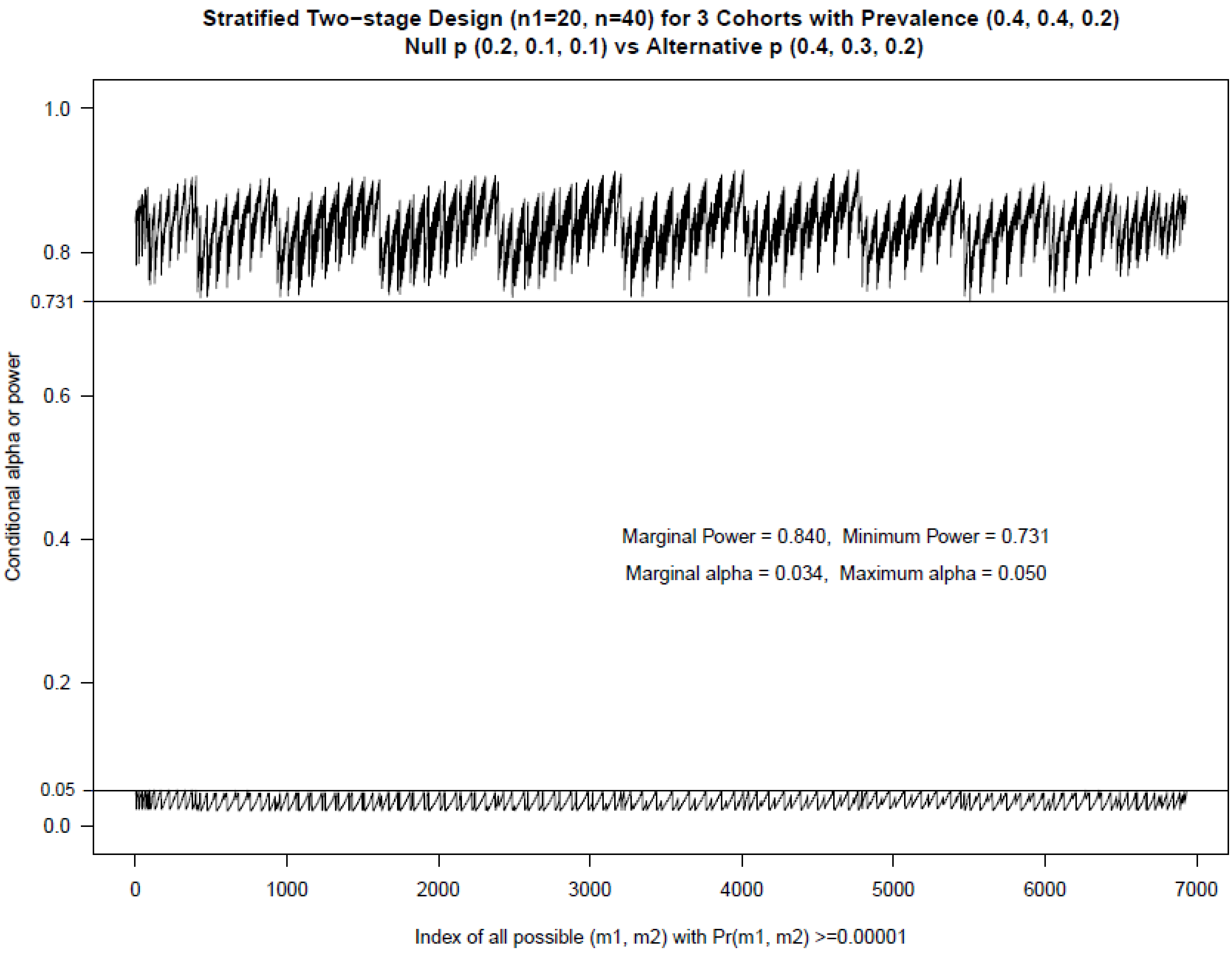
Threshold	Total Prob of cases with Pr(m1,m2)>=Threshold	Total # of cases with Pr(m1,m2)>=Threshold	Min Power
0.001	0.146	119	0.790
0.0001	0.808	2271	0.759
0.00001	0.975	6934	0.731
0.000001	0.997	12937	0.711



Stratified Two−stage Design (n1=20, n=40) for 3 Cohorts with Prevalence (0.4, 0.4, 0.2)
Null p (0.2, 0.1, 0.1) vs Alternative p (0.4, 0.3, 0.2)

Marginal Power = 0.840, Minimum Power = 0.476

Marginal alpha = 0.034, Maximum alpha = 0.050



Stratified Two−stage Design (n1=20, n=40) for 3 Cohorts with Prevalence (0.4, 0.4, 0.2)
Null p (0.2, 0.1, 0.1) vs Alternative p (0.4, 0.3, 0.2)

Marginal Power = 0.840, Minimum Power = 0.731

Marginal alpha = 0.034, Maximum alpha = 0.050

Single arm stratified designs

Strategies for testing multiple subsets, either ordered or unordered are also available (LeBlanc et al., 2012)

Eg, for a 2 cohort ordered subset design, one subset is expected to have better response to the new agent. Interim testing is done on this subset and if the alternative is rejected (ie, insufficient response) then the whole study is closed. If the study continues both the subset and overall group are tested after accrual is complete. Sample size is reduced vs testing each subset.

Single arm interpretation: marker studies

Early single arm trials are inadequate:

- If response rate is higher in marker positive than marker negative, unknown whether marker is prognostic or whether treatment is more effective in marker positive.

	Prognostic		Predictive of treatment effect	
	Marker +	Marker -	Marker +	Marker -
control	45%	15%	30%	30%
experimental	60%	30%	60%	30%

- Often limited or no historical information on response in the subsets
- Target may be poorly understood or poorly measured

Single arm interpretation: uncertainty about null

In a setting of less well characterized historical information, a three level decision rule might be useful. (Storer, 1992)

E.g., P_0 thought to be between P_{01} and P_{02}

Conclude promising if significantly better than P_{02}

Conclude unpromising if not significantly better than P_{01}

Otherwise additional phase II testing is needed.

If poorly characterized, consider randomized phase 2.

Single Arm Interpretation: use of early endpoints

Track record for single arm Phase II trials predicting Phase III success is poor.

Zia et al (2005)

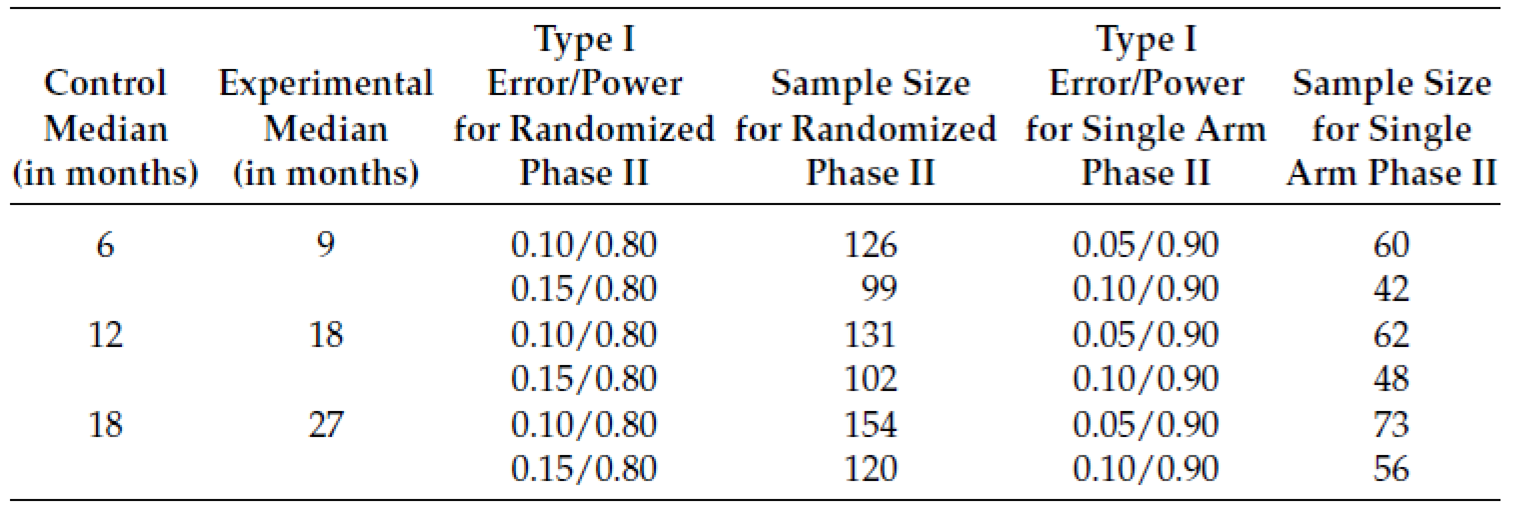
43 Phase III trials done after positive results in a Phase II with the same population and the same treatment

- In 35/43 response rates were lower in the Phase III
- In only 1 was the response rate substantially higher
- Only 12 of the phase IIIs were positive

Randomized Phase II

Randomized Phase II with control arm

If historical information is poorly characterized, a randomized Phase II using longer term endpoints and a control arm may be needed. Larger sample sizes are required.



Control Median (in months)	Experimental Median (in months)	Type I Error/Power for Randomized Phase II	Sample Size for Randomized Phase II	Type I Error/Power for Single Arm Phase II	Sample Size for Single Arm Phase II
6	9	0.10/0.80	126	0.05/0.90	60
		0.15/0.80	99	0.10/0.90	42
12	18	0.10/0.80	131	0.05/0.90	62
		0.15/0.80	102	0.10/0.90	48
18	27	0.10/0.80	154	0.05/0.90	73
		0.15/0.80	120	0.10/0.90	56

Phase 2 Selection Designs

Aim is to select which among several candidate regimens should be studied further.

Intent is not a definitive comparison, rather to choose for further study a treatment that is not likely much worse than the other candidate treatments. The arm observed best by any amount is chosen.

Sample size is chosen such that if one treatment is superior by Δ, then the probability of selecting it is π

Phase 2 Selection designs

Sample size for $\pi = .9$ (Simon et al., 1985; Liu et al., 1993)

	2 arms	3 arms	4 arms
Binomial			
$\Delta=.15$	37	55	67
Time to event			
HR=1.5	36	54	64

Allows for modest sample size to choose regimen for further study when not all can be taken to phase 3.

Limitation: A single arm is always chosen, even if none look useful, or if more than one looks useful. Addition of a threshold may be useful.

Phase 2 Randomized Discontinuation Design

Patients are all assigned with experimental treatment in a run-in phase to make the randomized group more homogeneous, allowing for fewer randomized patients

Patients who remain stable without serious toxicity at a specified time point are randomized between experimental treatment and either placebo or standard treatment

Sample size may be very large in order to randomize enough patients

Randomized discontinuation design

Limitation:

The question of primary interest – whether a phase III of up-front treatment with the new agent should be done may not be well addressed. The results may provide information on activity of the agent, but the clinical question in this design is whether patients with at least stable disease should continue with the new treatment.

May be useful in cases of potential cure: preliminary information on how often patients relapse after discontinuation

Proposed Risk Benefit Design

Motivation:

1. Pegilated liposomal doxorubicin above
2. Hematology agent

- 19% d/c due to AE.
- More more grade 3-4 and SAEs vs comparator
- Pts with reduced doses generally had good outcome.
- MTD lower in Phase 1 for different indications
- Missing assessment reduced ITT response estimates; other outcomes promising

Proposed Risk Benefit Design

- Interested in exploring lower doses,to identify a dose with potentially better benefit/risk profile
- Need efficient designs that assess both response and toxicity, i.e., benefit/risk trade off, with moderate sample size within budget

Work in collaboration with Tao Wang, Pfizer.

Risk Benefit Design

- Combined ideas from Thall and Cook (safety-efficacy scores describing risk benefit trade off) and selection designs from Simon.

- Select the best treatment from a set of candidates
  - Ensure adequate probability of correct selection if an arm has benefit/risk that is acceptable and superior to other arms by Δ
  - Control probability of selecting an arm with unacceptable benefit/risk

Risk Benefit Design

Step 1: Define benefit/risk score and acceptable benefit/risk region.

- Score is $S(p_R, p_D) = D(\mathbf{p}) = 1 - \left\{ \left(\frac{p_d - 1}{\lambda_1 - 1} \right)^{\beta} + \left(\frac{p_r}{\lambda_2} \right)^{\beta} \right\}^{1/\beta}$

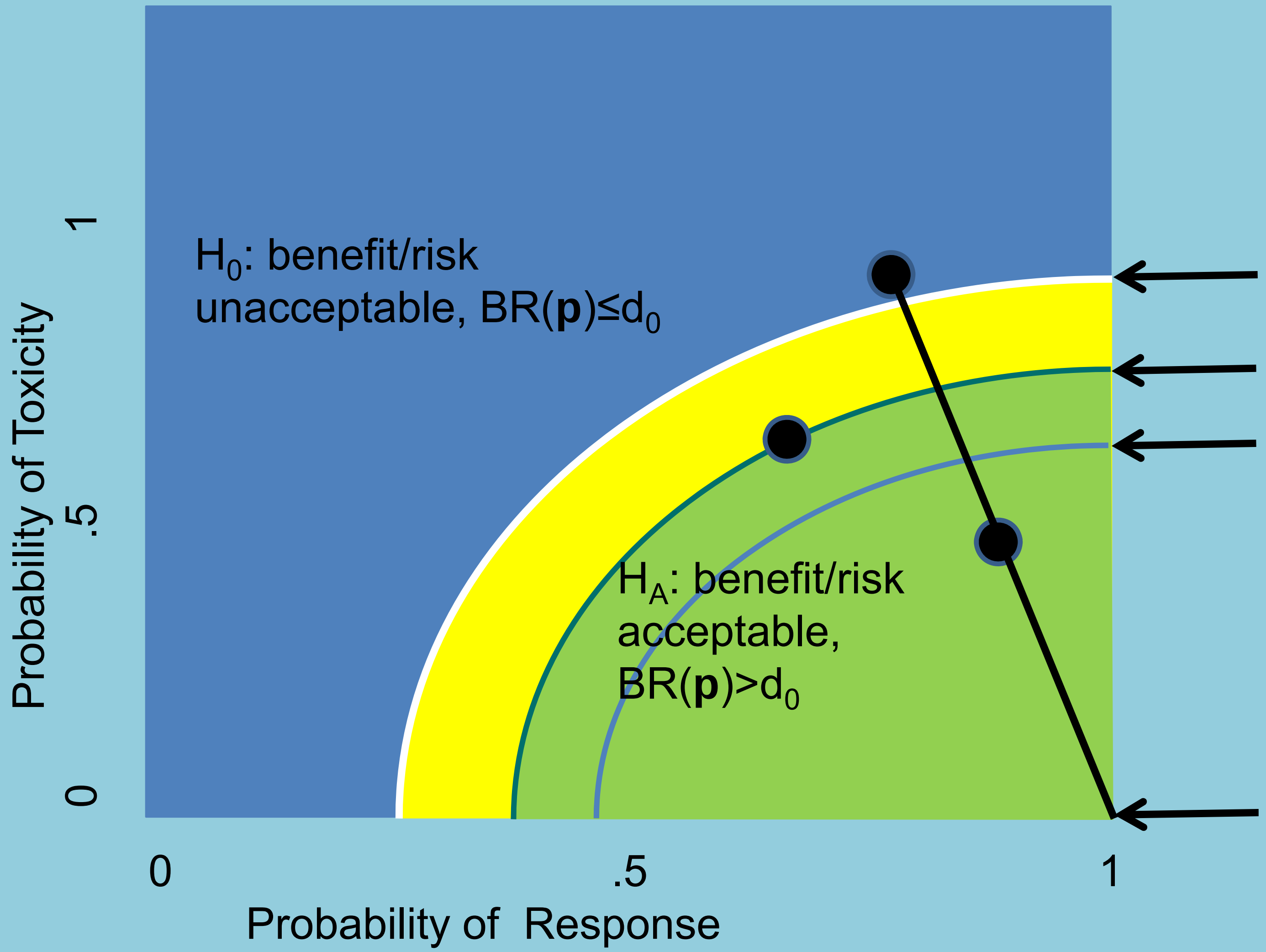
  with contours determined by choice of 3 points with equivalent benefit risk (Thall & Cook)

Step 2: For each dose test for acceptable benefit/risk

Step 3: Select best dose from acceptable dose(s) from step 2

- If more than one dose is acceptable, choose the one with observed best result

Decision Framework: Null and Alternative in Step 2


H_0: benefit/risk unacceptable, BR(**p**)≤d_0
H_A: benefit/risk acceptable, BR(**p**)>d_0
1
.5
0
Probability of Toxicity
0
.5
1
Probability of Response

d_0, null contour

D, cutoff value

d_A, alt contour

Ideal of 100% efficacy and 0% toxicity

p is the true (p_r, p_d) for a treatment arm

Risk Benefit Design

Testing acceptability

- Contours for the hematology example were determined by specification of (.85,.2), (.80,.16), (.75, .10) as being of = BR
- Decision criteria chosen such that probability of selecting a dose in the unacceptable region is reasonably low:
- Reject H_0 if $BR(\mathbf{X}) \geq D$, where D satisfies $\sup \{\inf [d: P(BR(X) \geq d \mid \mathbf{p},\Phi] \mid \mathbf{p}\ \varepsilon\ H_0 \} \leq \alpha$ where X=, $(\hat{p}_{R/NoD}, \hat{p}_{NoR/NoD}, \hat{p}_{R/D}\ , \hat{p}_{NoR/D})$ $\mathbf{p}=(p_R, p_D)$ and Φ is the odds ratio

Risk benefit design

Power for a contour

- For a specific contour in the alternative region (H_A), C, with BR= BR(C), the point on the contour with the smallest probability of rejecting H_0 defines the power
- Power ≥ inf [P(BR(X) ≥ D | **p,** Φ) | **p** ε C]

Risk benefit design

Determination of D and power

- For a given ($\mathbf{p}, \phi$), exact enumeration was used to determine the cutoff and power
- For an arm with n=100 patients, there are 176851 possible 2x2 tables $\mathbf{X}=(X_{11}, X_{10}, X_{01}, X_{00})$ in total
  - 10198 unique values of $BR(\mathbf{X}|\mathbf{p}, \phi)$

Risk benefit design

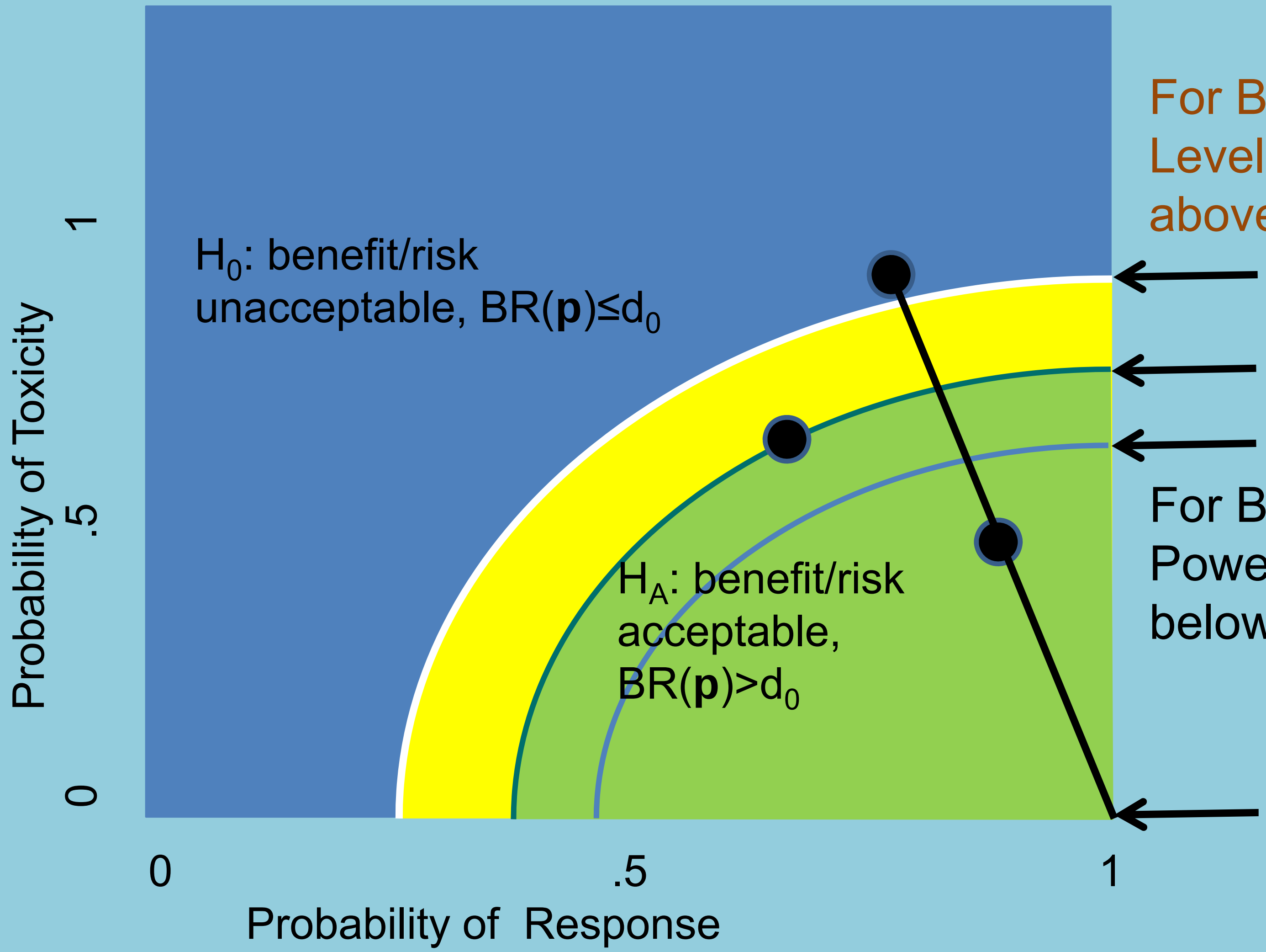
Determination of D and power cont.

- Distribution of test statistics BR($\mathbf{X}|\mathbf{p},\Phi$) can be derived accordingly P(BR($\mathbf{X}|\mathbf{p},f$)=d) = Σ_i {P($\mathbf{X}_i|\mathbf{p},\Phi$): all i ε [1,…,176851] such that BR($\mathbf{X}_i$)=d}
- Minimization and maximization of P(BR($\mathbf{X}$) ≥ D | **p**, Φ) over a contour - use R function to optimize (golden section search and successive parabolic interpolation) or nlminb

Risk benefit design

- After evaluating different options for the hematology example, BR = -0.4 was chosen as the null contour, BR = 0.1 for the alternative contour, and D was determined to be -0.085 for a sample size of 100.
- The cutoff and power are fairly insensitive to assumptions about ϕ; the simplified assumption of independence was used. Similar findings were reported by Bryant and Day (1995) and Conaway and Petroni (1996)

Decision Framework: Null and Alternative in Step 2


H_0: benefit/risk unacceptable, BR($\mathbf{p}$)≤d_0
H_A: benefit/risk acceptable, BR($\mathbf{p}$)>d_0
Probability of Toxicity
0
.5
1
0
.5
1
Probability of Response

For BR = d_0, Level ≤.056 on or above this curve

d_0, null contour

D, cutoff value

d_A, alt contour

For BR = d_A, Power ≥85% on or below this curve

Ideal of 100% efficacy and 0% toxicity

p is the true (p_r, p_t) for a treatment arm

Risk benefit design

– D was determined to be -0.085

Examples of Min and Max Power for given contours

BR(C)	**Min Power**	**Max Power**
-0.4 (null)	0.0244	**0.0558**
-0.2	0.216	0.296
-0.085 (D)	0.474	0.494
0	**0.687**	0.696
0.1 (alternative)	**0.856**	0.883

Risk benefit design

- Selection Power: Probability of choosing Arm A when, in fact, Arm A is acceptable and superior by a specified amount over other arms
- Selection Type I Error Rate: Probability of selecting an arm with unacceptable benefit/risk

P of Selecting Arm A among 2 Arms

- Minimum and maximum probability of selecting arm A with (p_r, p_t) on contour A with BR=BR$_A$, over Arm B with (p_r, p_t) on contour B with BR=BR$_B$

Arm B BR	Arm A BR					
	-0.4 (d_0)	-0.2	-0.085	0	0.1 (d_A)	0.2
-0.5	0.0243 0.0557	0.215 0.294	0.472 0.540	0.685 0.719	0.854 0.882	0.955 0.972
-0.4 (d_0)	0.0238 **0.0551**	0.211 0.293	0.466 0.537	**0.668** 0.718	**0.8495** 0.880	0.952 0.970
-0.2	0.0201 0.0294	0.184 0.264	0.417 0.473	0.621 0.668	0.800 0.844	0.919 0.952
0	0.0108 0.0251	0.109 0.158	0.267 0.327	0.430 0.483	0.614 0.670	0.773 0.823
0.1 (d_A)	0.0055 0.0145	0.0643 0.098	0.170 0.209	0.293 0.335	0.460 0.512	0.633 0.692

Increasing

Increasing

P Selecting Arm A among 3 Arms

- Minimum and maximum probability of selecting arm A with (p_r, p_t) on contour A with BR=BR_A, over Arm B with (p_r, p_t) on contour B with BR=BR_B and Arm C with (p_r, p_t) on contour C with BR=BR_C

Arm B, C BR	**Arm A BR**					
	-0.4 (d_0)	-0.2	-0.085	0	0.1 (d_A)	0.2
-0.5, -0.5	0.0241 0.0551	0.213 0.295	0.470 0.492	0.672 0.694	0.853 0.882	0.954 0.972
-0.4, -0.4	0.0232 **0.0544**	0.207 0.290	0.459 0.487	**0.661** 0.689	**0.843** 0.878	0.905 0.970
-0.2, -0.2	0.0167 0.0417	0.158 0.237	0.369 0.454	0.560 0.607	0.750 0.816	0.886 0.933
-0.4, 0	0.0106 0.0254	0.107 0.158	0.264 0.325	0.426 0.479	0.610 0.673	0.770 0.832

Increasing

Increasing

Benefit risk design: Summary

The proposed Benefit/Risk design addresses objectives

- Moderate sample size
- Selecting most promising dose based on benefit/risk trade off vs informal assessment
- Controlling the probability of selecting an arm with unacceptable risk benefit (selection type I error)
- Adequate selection power to choose the correct treatment arm when one arm is superior by a specified amount to the other arms

Benefit risk design: Summary

Limitations

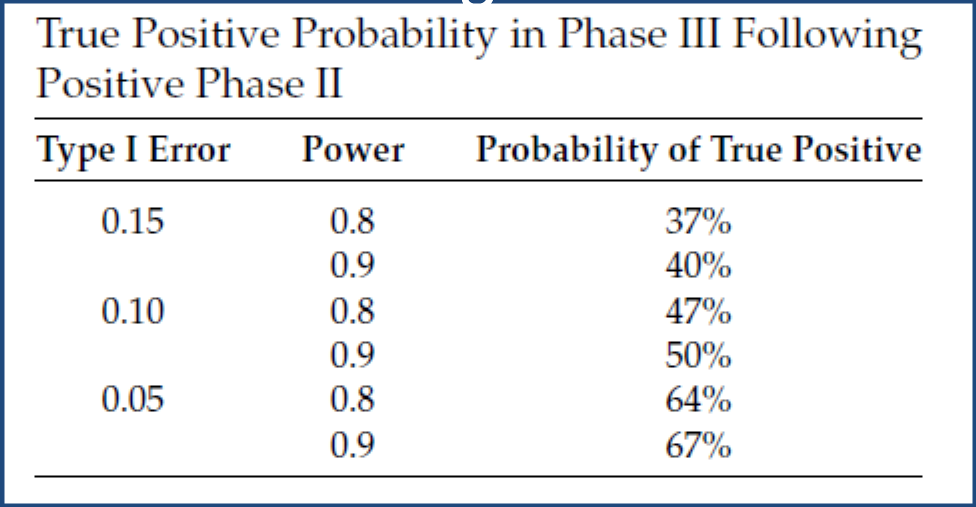
- Computing intensive
- A larger study may be needed to confirm the selected dose
- Subjectivity in determining benefit/risk score
  - Advice needed

Phase 2 strategy

As noted before, track record for phase 2 prediction of phase 3 success is poor.

Part of the reason is that many new agents are not effective. The % of positive results taken to Phase III that are true positives may be quite low.

Assuming 10% of new agents are active:



True Positive Probability in Phase III Following Positive Phase II

Type I Error	Power	Probability of True Positive
0.15	0.8	37%
	0.9	40%
0.10	0.8	47%
	0.9	50%
0.05	0.8	64%
	0.9	67%

Phase 2 strategy

Consider various Phase II strategies:

Single arm, one-sided level .05, power .9 for testing H_0: $p_E \leq .25$ vs. H_A: $p_E > .45$, sample size 50

Single arm, three level: go to Phase III if significantly >.35; discontinue if not significantly >.25; randomized Phase II if neither. Sample size 50 for first trial.

Phase 2 strategy

Adaptive approach

- Randomize and accrue 50 patients to E and 25 to C.
- C is used to update the historic control estimate, h_0. k/25 responses are observed. h_0 is updated to $P_0 = .5(h_0) + .5(k/25)$.
- Go/No Go decision rule for E is based on single arm test of P_0 vs $P_0 + .2$.
- Sometimes power is good, sometimes level

Phase 2 Strategy

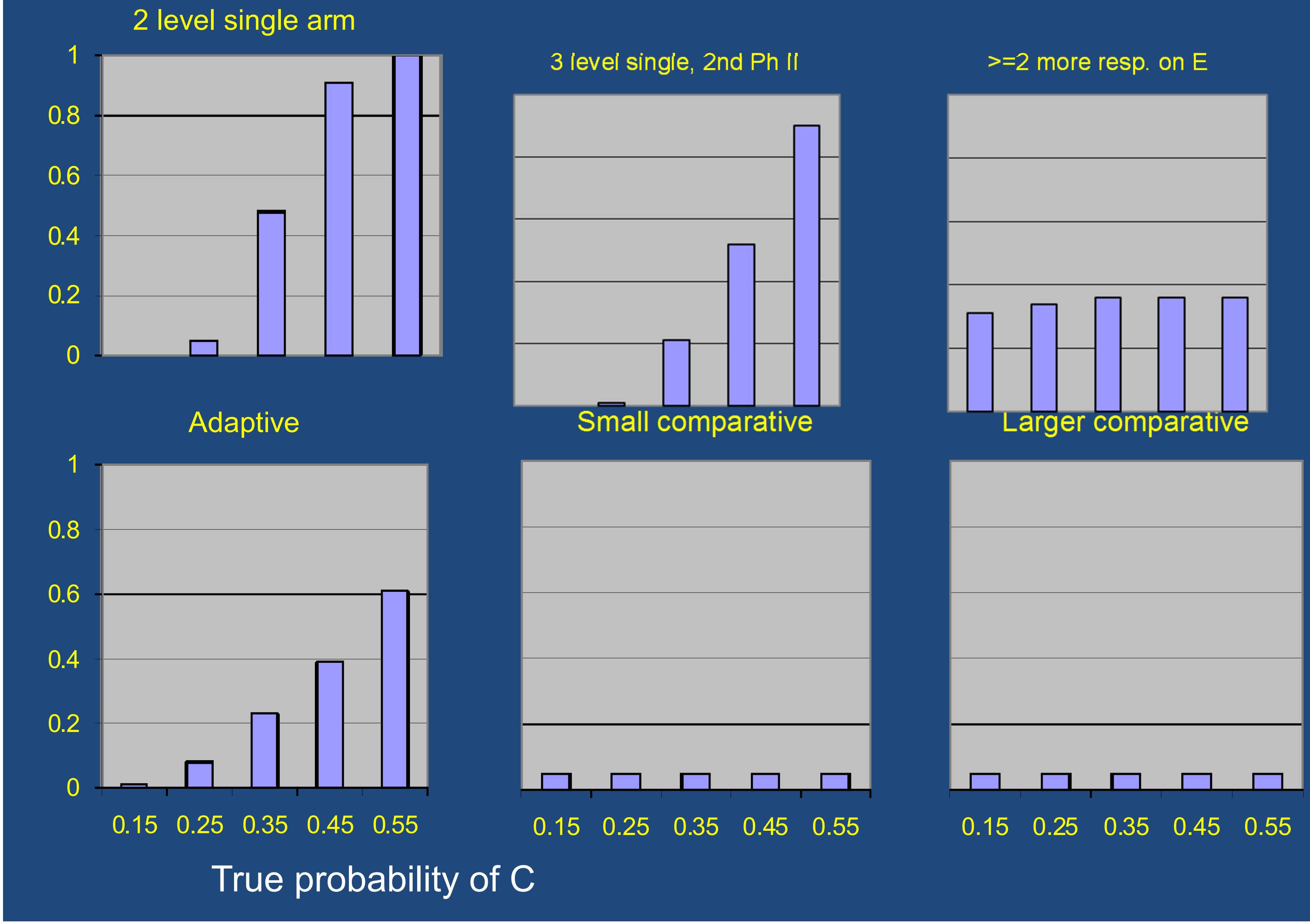
Randomized:

Good Level, small sample size: Test H_0: $p_E \leq p_C$ vs. H_A: $p_E > p_C$ with a conventional test of one-sided level .05 and sample size 74

Good power, small sample size: Modified selection design. Go to Phase III if at least X more successes on the experimental arm than on control. Sample size 74.

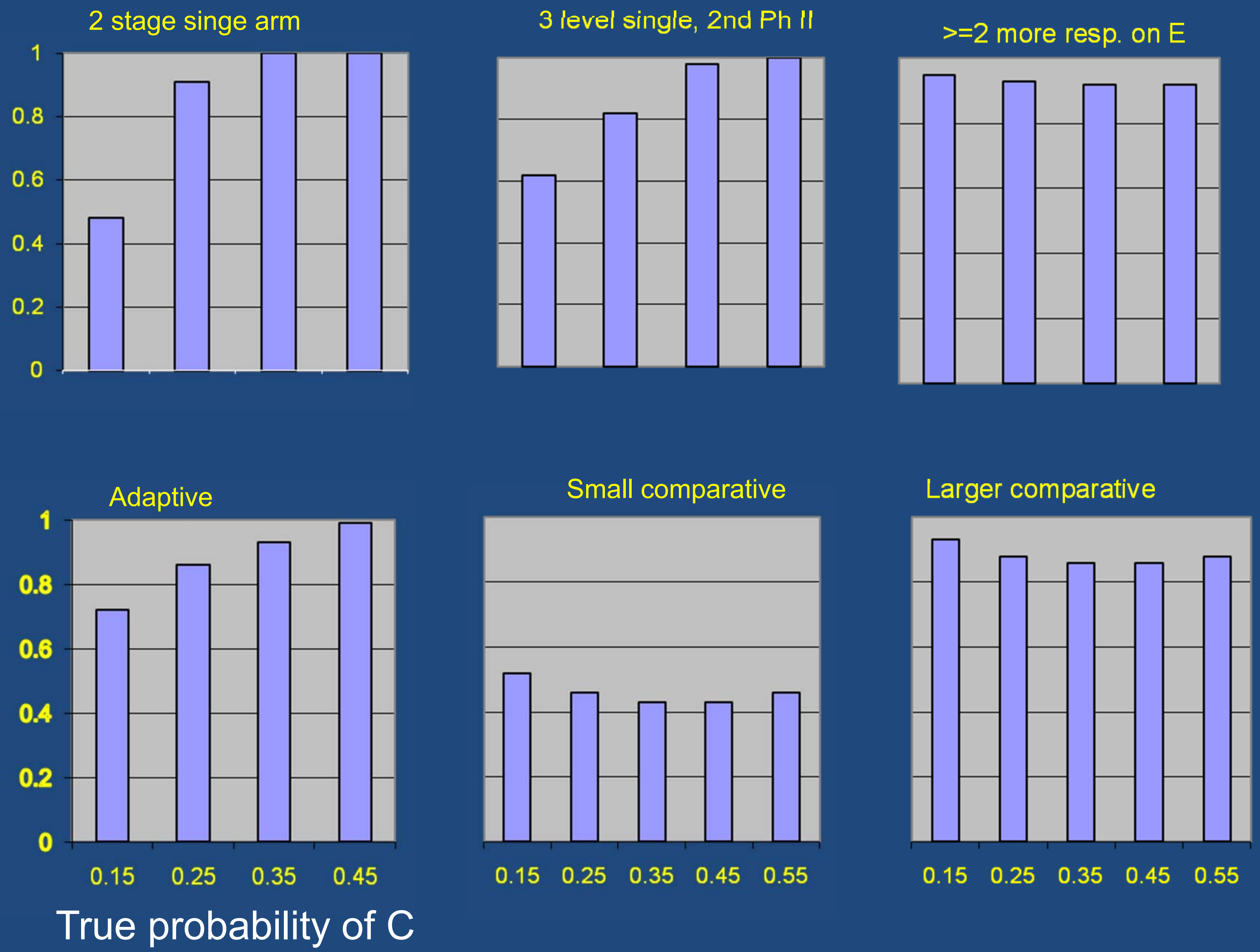
Good Level, good power, larger sample size: Test H_0: $p_E \leq p_C$ vs. H_A: $fp_E > p_C$ with a conventional test of one-sided level .05 and sample size 200

Comparison: Probability of positive result when C=E



E = C for all of these cases

Comparison: Probability of positive result when E>C



E = C + 0.2 for all of these cases

Phase 2 strategy

Ultimately we want the phase 2 strategy to result in good probability of Phase 3 success

Illustration:

If we assume estimates are fairly good with

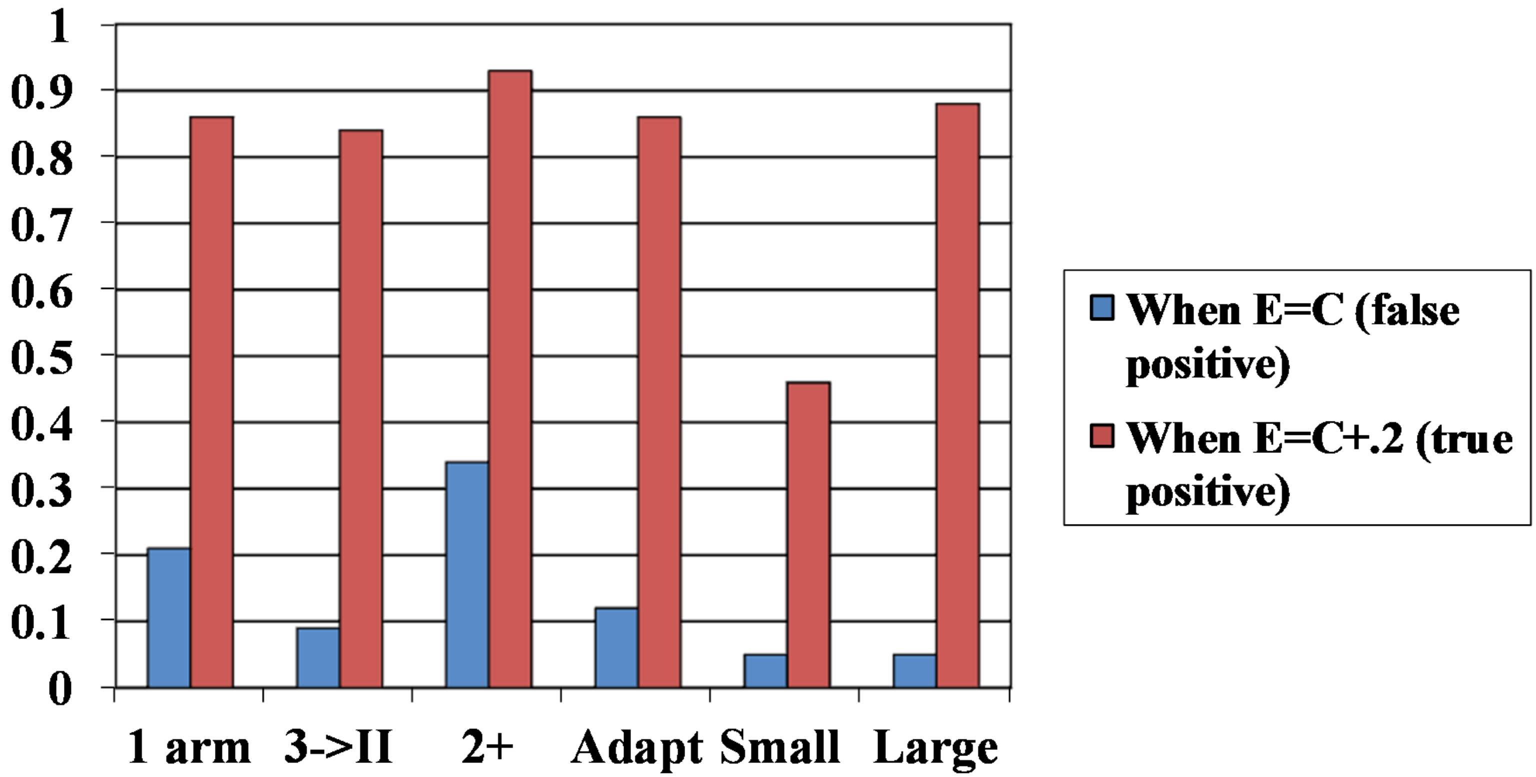
P(historical assumption correct)=.4

P(historical is .1 lower than assumed)=.2

and P(historical is .1 higher than assumed)=.4…

Phase 2 strategy, example

Then probabilities of phase 2 success (and phase 3 development) will be



Phase 2 strategy, example

Expectations of activity of new agents:

Assume only 20% of Phase IIs show activity

Assume half of agents active with respect to response will result in survival improvement

Assume a few “inactive” agents will result in survival improvement --

Then in 100 phase 2s it is expected that 14 of the regimens studied have survival benefit.

If 100 phase 2s are done and positives only go to Phase 3, then..

100 phase 2s. 14 potential positive Phase IIIs.

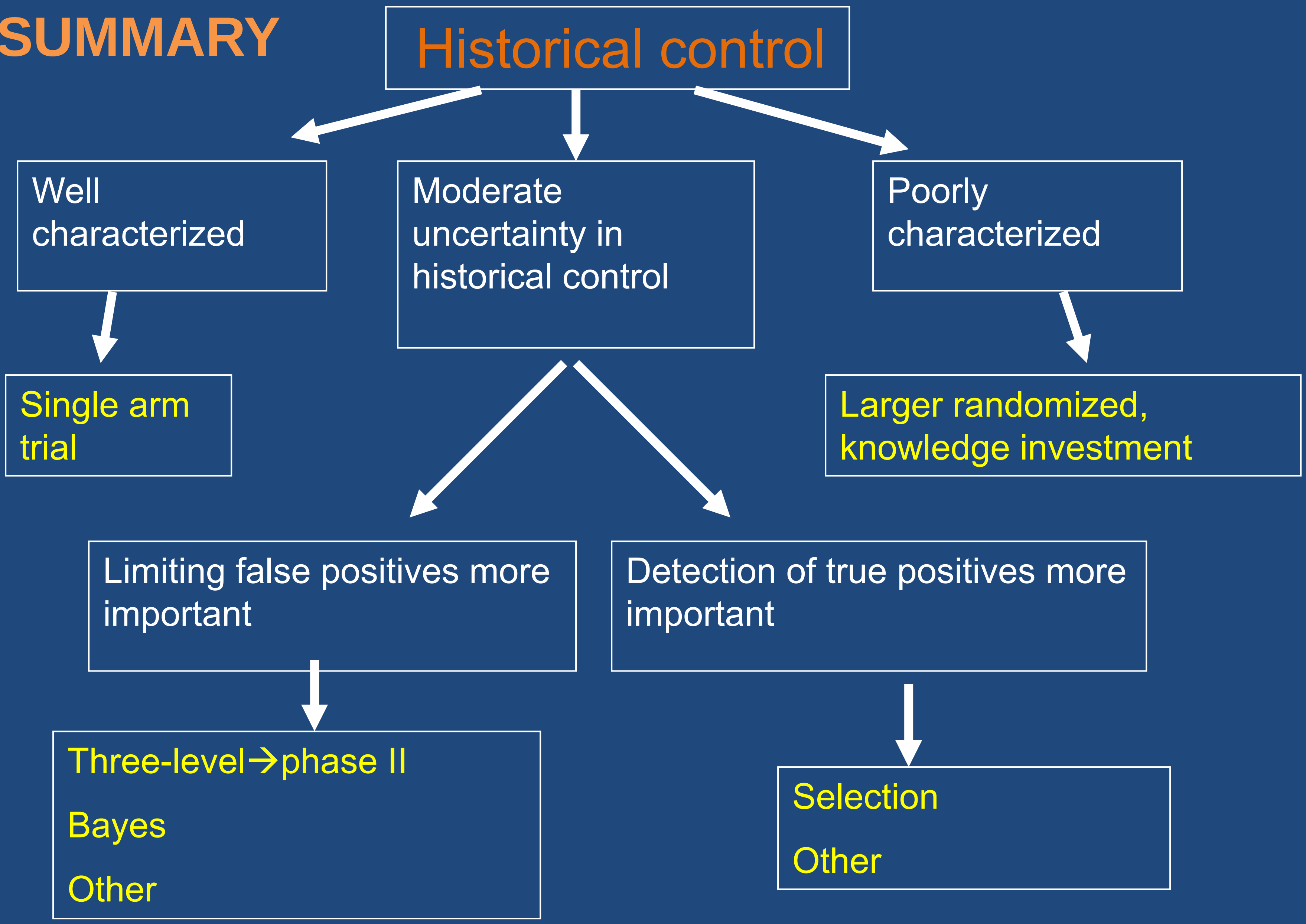
Strategy	No. Ph 2 positive= no. Ph 3 done	No. Ph 3 pos.	% of potential 14 pos.	% Phase IIIs that are pos.
Single arm	34	9	64%	26%*
Three level->Ph II	24	9	64%	38%
Modified Selection	46	11	79%	24%
Adaptive	27	9	64%	33%
Small comparative	13	5	36%	38%
Large comparative	22	9	64%	41%

* Similar to Zia, 12/43

Phase 2 strategy trade-offs

Strategy		% of 14	% + Ph IIIs
Single arm	Cheapest Ph IIs	64%	26%
Three level ->Ph II	Expense, delayed decision	64%	38%
Selection	Low Ph III yield	79%	24%
Adaptive	No delay	64%	33%
Small comparative	Missed active agents	36%	38%
Large comparative	100 Expensive Phase IIs	64%	41%

SUMMARY



Historical control

Well characterized

Moderate uncertainty in historical control

Poorly characterized

Single arm trial

Larger randomized, knowledge investment

Limiting false positives more important

Detection of true positives more important

Three-level→phase II

Bayes

Other

Selection

Other

Phase 2 Conclusions

- Single arm Phase IIs may be useful when historical results are stable and well characterized
- Consider alternative designs, eg, when there are subset issues, or when formal risk-benefit assessment is an objective.
- Reliability of phase 2 results is limited but can be improved.
- Phase 3 success rate can be improved modestly by improved Phase II strategy

Optimal design

- Optimality can be relative only to limited criteria, but every improvement helps..