Adaptive Biomarker Trial Designs

Adrian Mander

MRC Biostatistics Unit, University of Cambridge

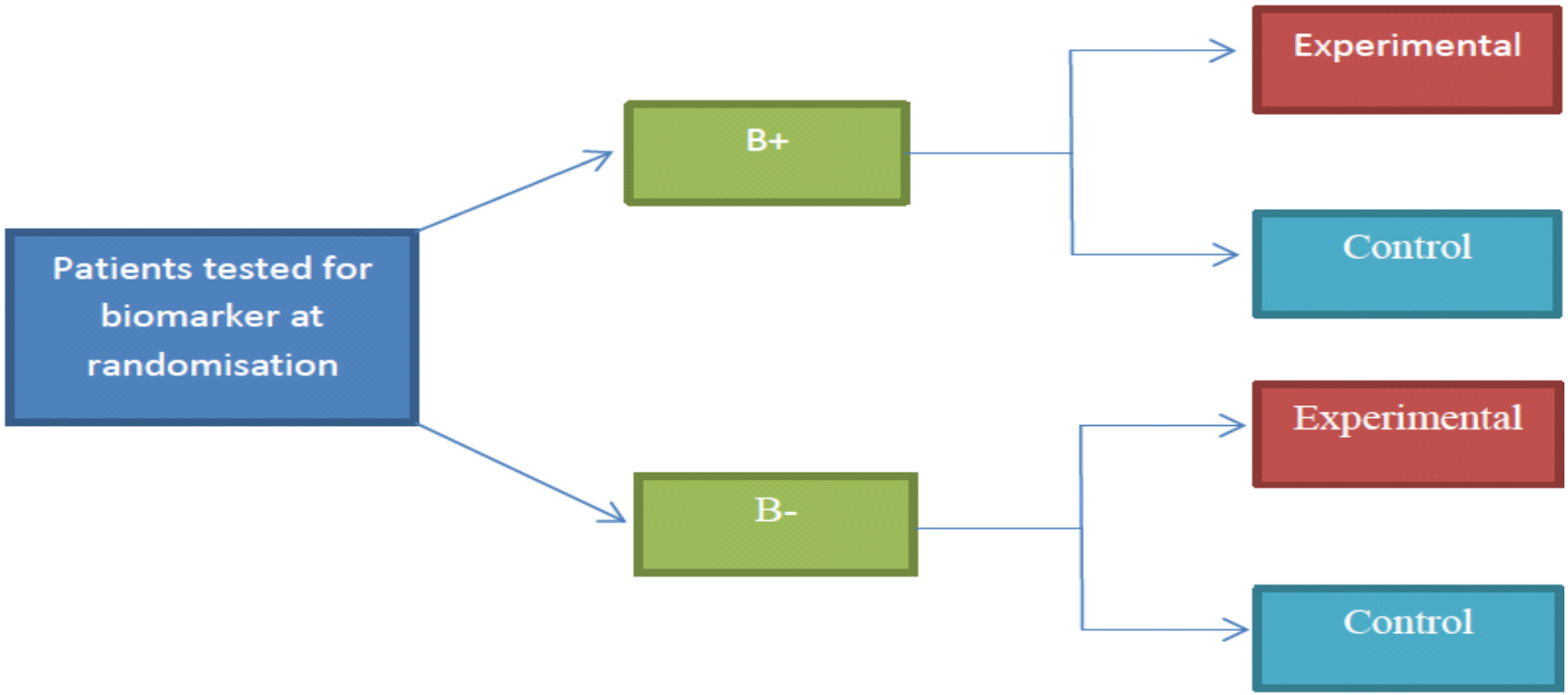
Jul 2017

Overview

1. Designs
2. Adaptive enrichment single-arm trial
3. Multiple treatments

Designs for single experimental treatment

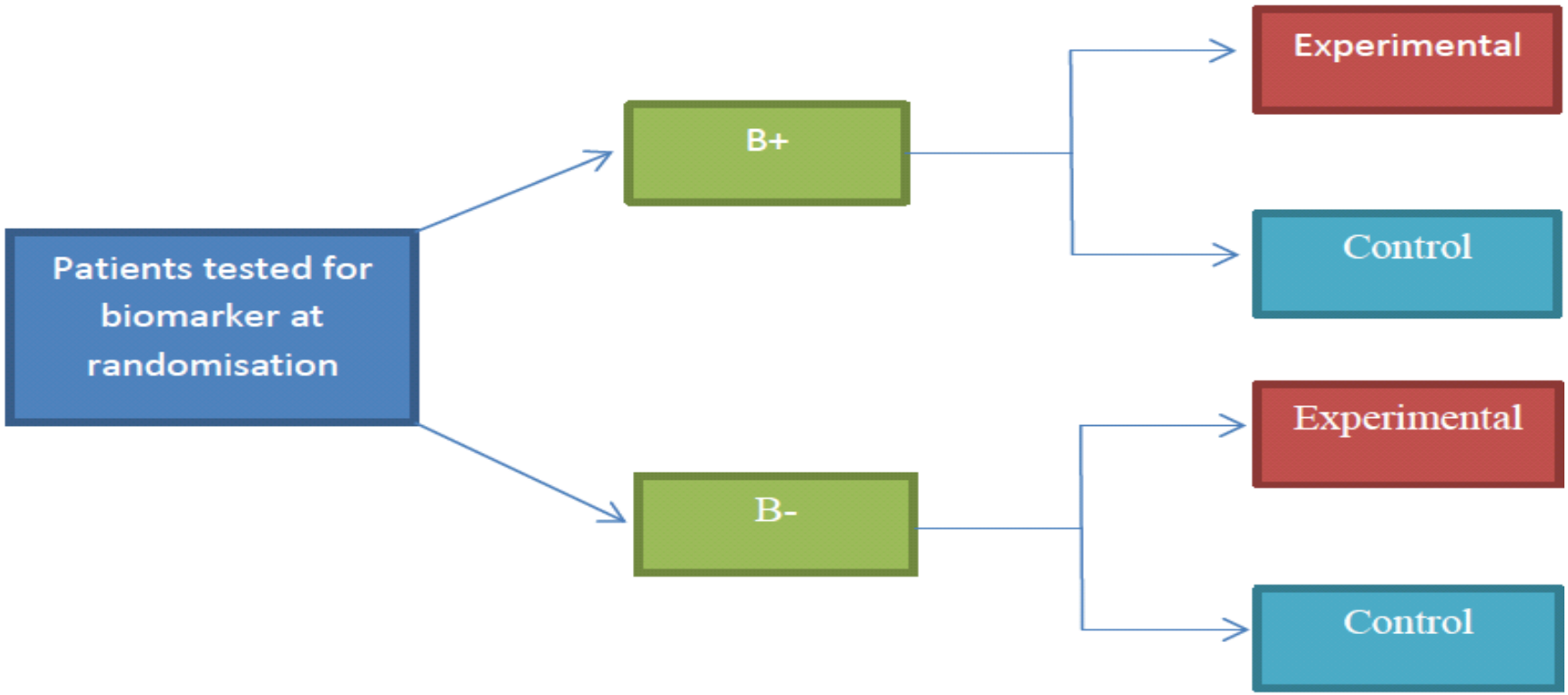
- **Marker by treatment** design; similar to a traditional RCT.
- Can test effect of experimental versus control, and whether B is a predictive biomarker.



Better than retrospective analysis should have higher power

Designs for single experimental treatment

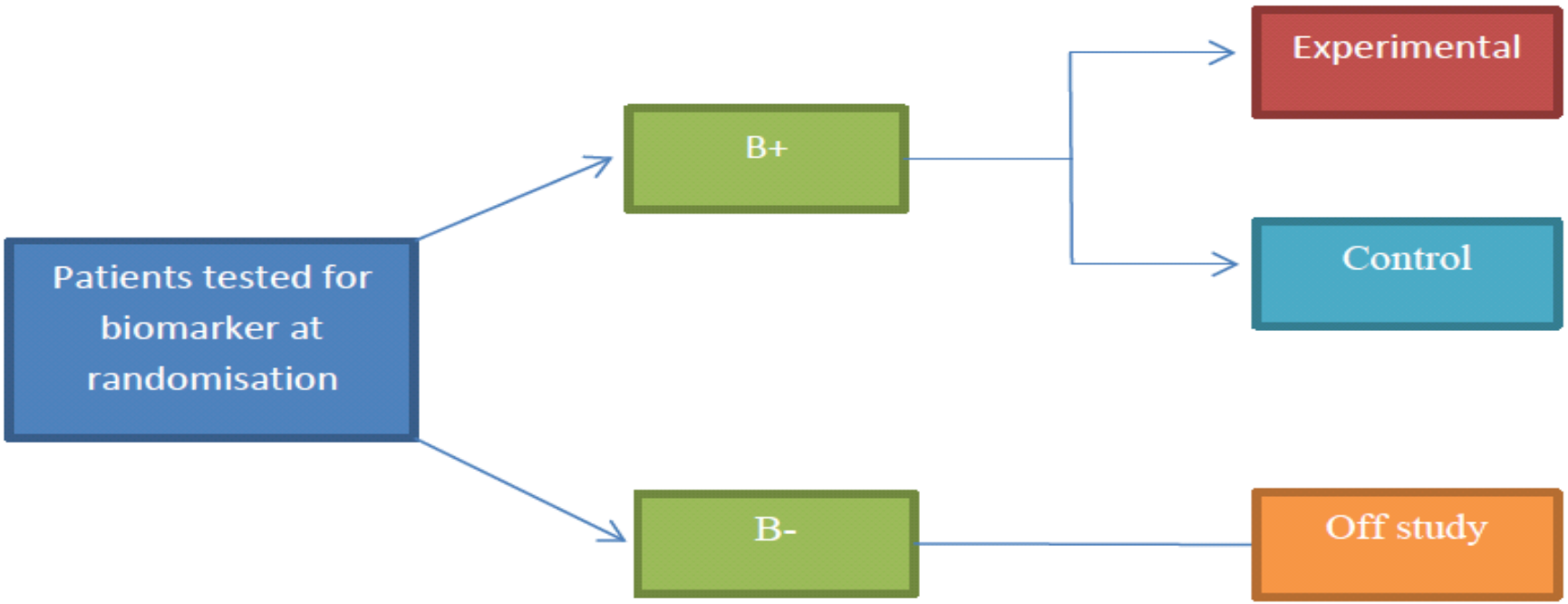
- **Marker by treatment** design; similar to a traditional RCT.
- Can test effect of experimental versus control, and whether B is a predictive biomarker.



Better than retrospective analysis should have higher power

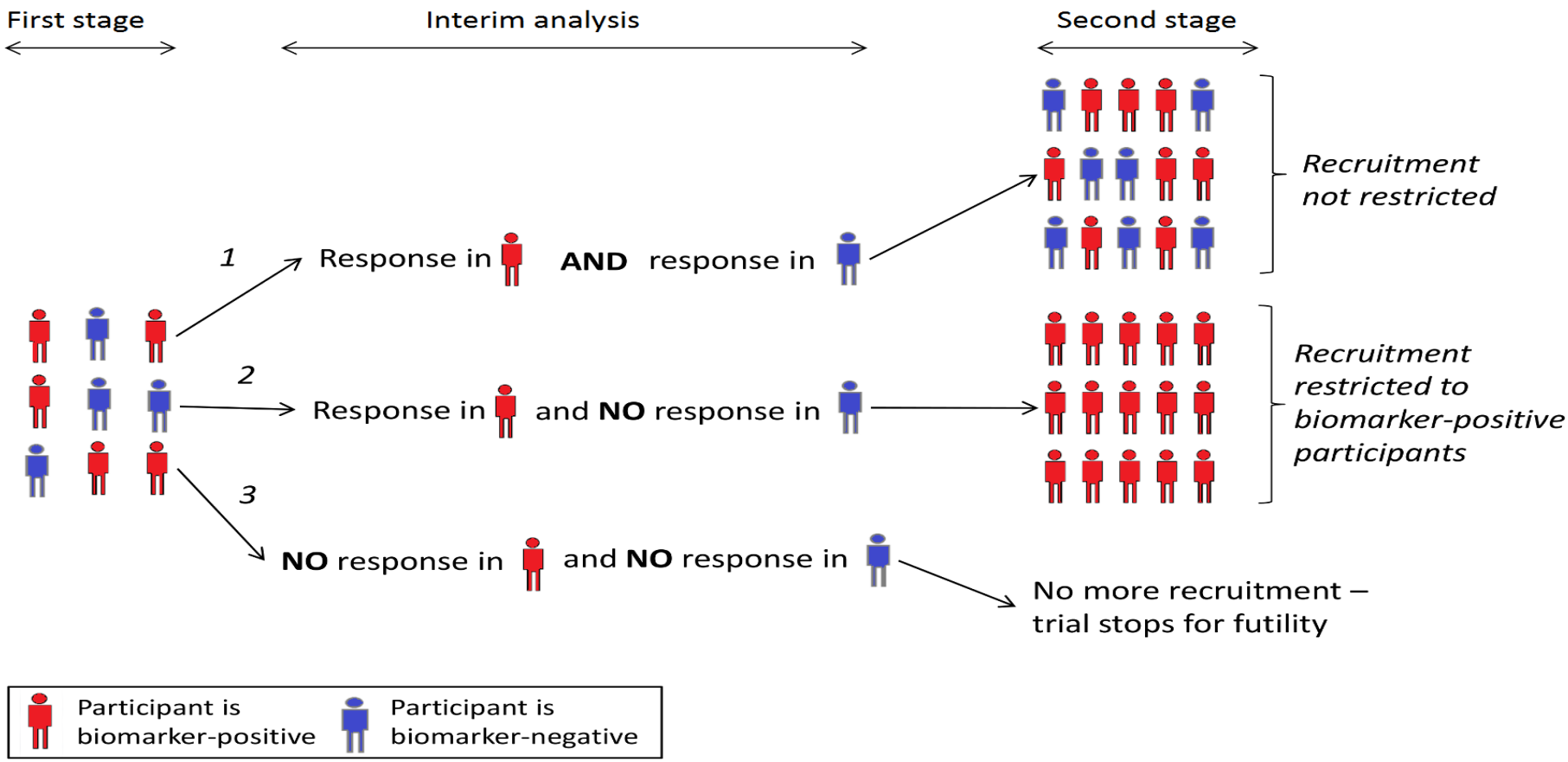
Enrichment

- **Enrichment design**: used if effect is likely only in B+ group.
- Assume biomarker is predictive



Adaptive Enrichment

- **Adaptive enrichment designs** recruit all patients, then have interim analysis to decide if recruitment should be restricted.
- Has a chance to find treatment effect in everyone if it is present, or to enrich if not.



Adaptive Enrichment Single-arm Trials

- Describe an adaptive enrichment study by Jones and Holmgren[1]
  - Revise single-arm designs and error calculations
  - Cover hypothesis testing (not covered by J&H well)
  - Finding optimal designs (minimise expected sample size)

[1] CL Jones and E Holmgren (2007) Clin Trials. 28(5):654-61. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies

Jones and Holmgren

Aim to design a trial for a targeted cancer therapy, made possible with improvements in molecular/genetic characterisation of biological pathways

- Outcome is (tumour) response/activity (RECIST)
- Determine whether drug has activity only in target population or as a whole
- Single-arm trial
  - powerful small study although sample sizes approach Phase III setting in the biomarker setting
- They base their design on Simon two-stage and introduce adaptive enrichment

Downsides

- Population selection bias of a one-armed trial!
- Single-arm trials have limited usefulness [1]

[1] MJ Grayling and AP Mander (2016) Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharmaceutical statistics 15 (2), 143-151

Jones and Holmgren

Aim to design a trial for a targeted cancer therapy, made possible with improvements in molecular/genetic characterisation of biological pathways

- Outcome is (tumour) response/activity (RECIST)
- Determine whether drug has activity only in target population or as a whole
- Single-arm trial
  - powerful small study although sample sizes approach Phase III setting in the biomarker setting
- They base their design on Simon two-stage and introduce adaptive enrichment

Downsides

- Population selection bias of a one-armed trial!
- Single-arm trials have limited usefulness [1]

[1] MJ Grayling and AP Mander (2016) Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharmaceutical statistics 15 (2), 143-151

Jones and Holmgren

Aim to design a trial for a targeted cancer therapy, made possible with improvements in molecular/genetic characterisation of biological pathways

- Outcome is (tumour) response/activity (RECIST)
- Determine whether drug has activity only in target population or as a whole
- Single-arm trial
  - powerful small study although sample sizes approach Phase III setting in the biomarker setting
- They base their design on Simon two-stage and introduce adaptive enrichment

Downsides

- Population selection bias of a one-armed trial!
- Single-arm trials have limited usefulness [1]

[1] MJ Grayling and AP Mander (2016) Do single-arm trials have a role in drug development plans incorporating randomised trials? Pharmaceutical statistics 15 (2), 143-151

Simon two-stage design - a recap

$$\text{Testing} \qquad H_0 : p = p_0$$

- Set the design parameters for a particular trial
  - 5% significance, 80% power
  - the null response of 5% and power at a response of 25%
- Discover optimal design is $0/12 \;\; 2/16$
  - in first stage: stop for futility if $0/12$ responders
  - at end of trial: reject H_0 if $> 2/16$ responders

The probability of rejecting H_0 in terms of p the response

1 - Probability of NOT rejecting H_0

$$= 1 - (B(12, 0, p) + b(12, 1, p) * B(4, 1, p) + b(12, 2, p) * B(4, 0, p))$$

$B()$ is $P(X \leq x)$, $b()$ is $P(X = x)$ and X is a Binomial distribution

Simon two-stage design - a recap

$$\text{Testing} \qquad H_0 : p = p_0$$

- Set the design parameters for a particular trial
  - 5% significance, 80% power
  - the null response of 5% and power at a response of 25%
- Discover optimal design is $\;0/12\;\;2/16$
  - in first stage: stop for futility if $0/12$ responders
  - at end of trial: reject H_0 if $> 2/16$ responders

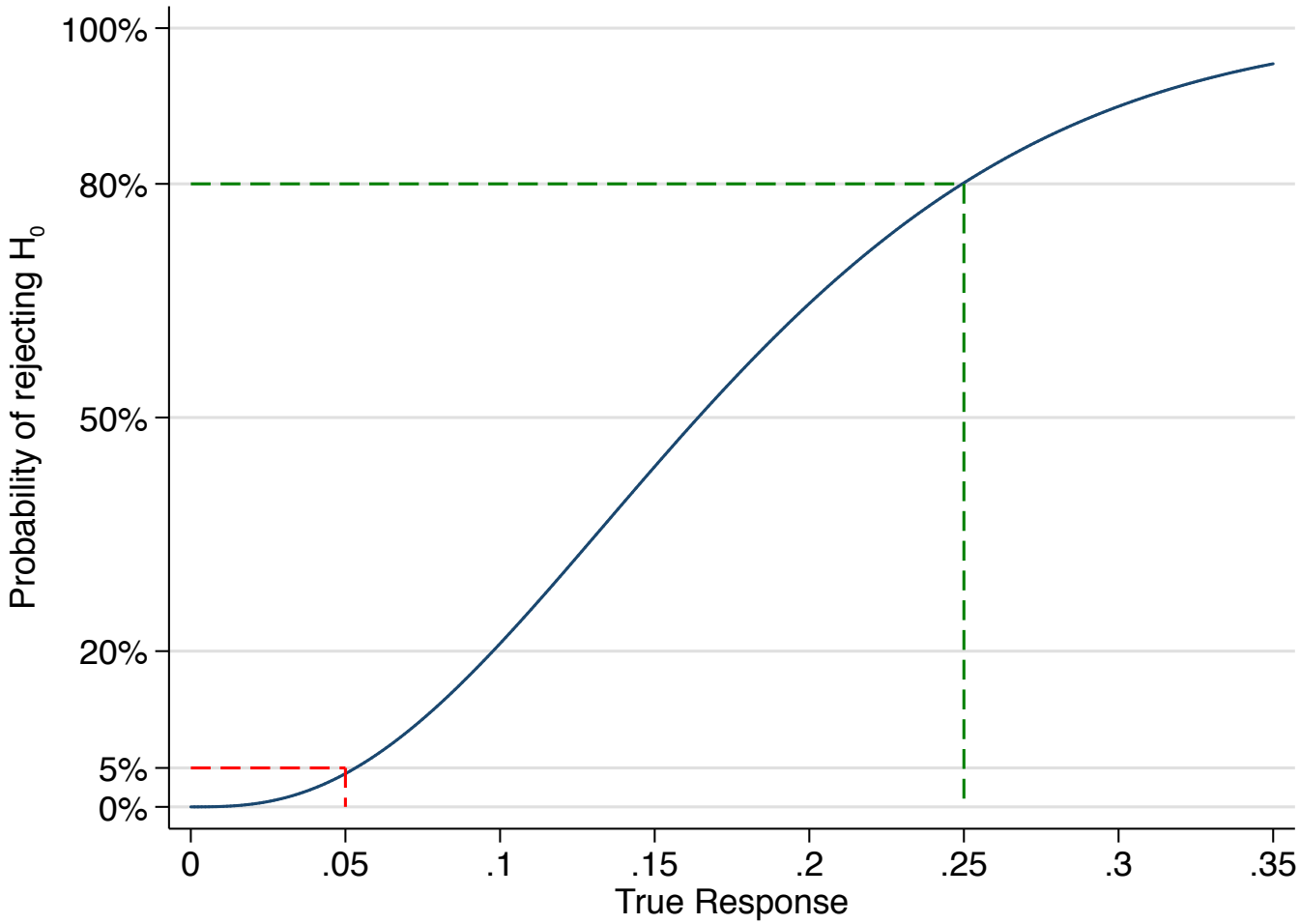
The probability of rejecting H_0 in terms of p the response

1 - Probability of NOT rejecting H_0

$$= 1 - (B(12, 0, p) + b(12, 1, p) * B(4, 1, p) + b(12, 2, p) * B(4, 0, p))$$

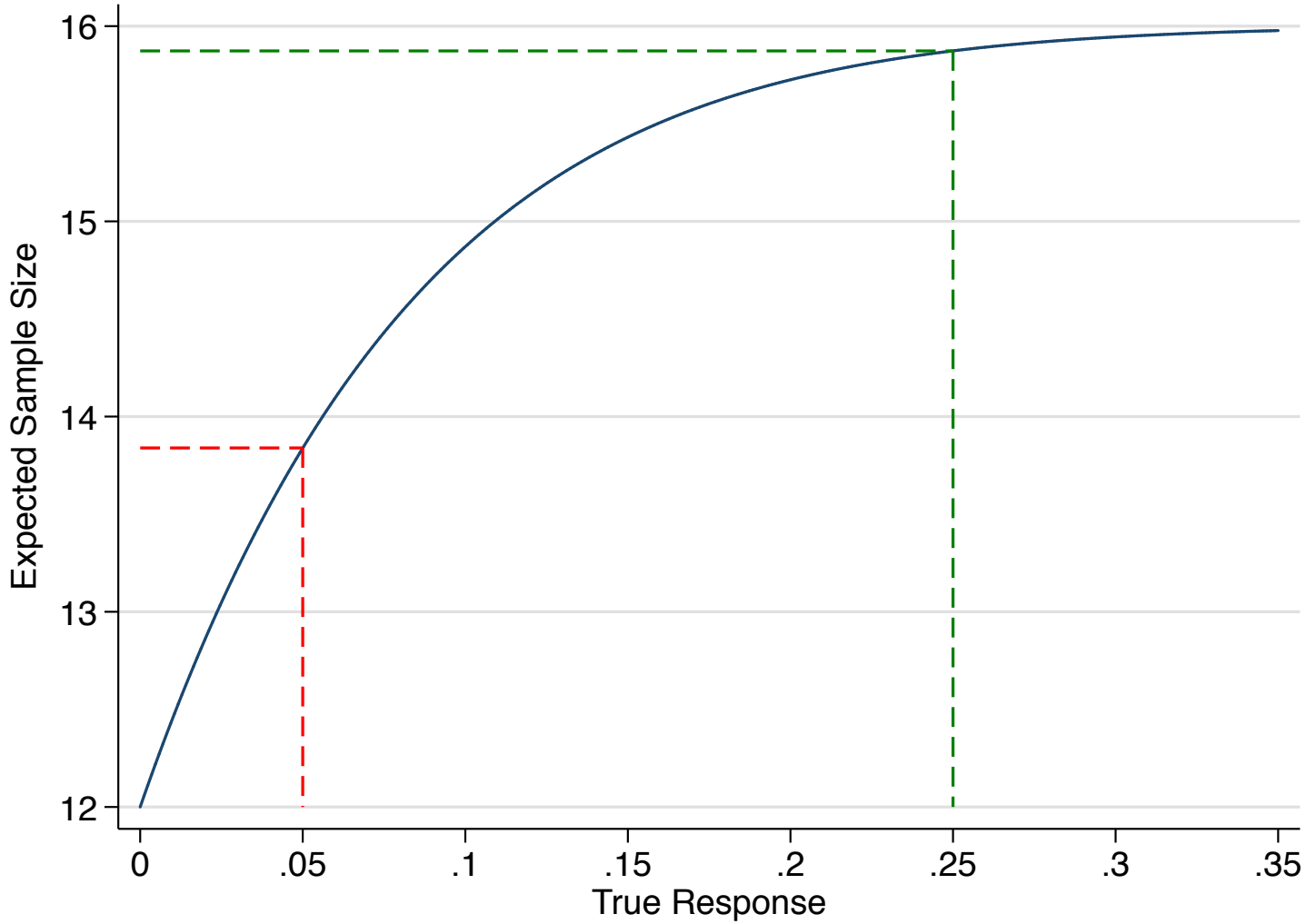
$B()$ is $P(X \leq x)$, $b()$ is $P(X = x)$ and X is a Binomial distribution

Probability of rejecting H_0



We have control of 5% significance if $p \leq 0.05$ and 80% power if $p \geq 0.25$ (Monotonicity allows us to write an inequality in the null hypothesis)

The expected sample size of the trial



Jones and Holmgren Design

Tests the **two** null hypotheses for the positive and the unselected population

$$H_0^- : p^- = p_0 \quad \& \quad H_0^+ : p^+ = p_0$$

- If you reject H_0^-
  - Conclude efficacy in unselected population
- If you reject H_0^+
  - Conclude efficacy in biomarker positive population

They assume that the response $p^+ > p^-$

Our design parameters

- $p_0 = 0.05$ (under null biomarker is not prognostic)
- 5% significance and 80% power

Jones and Holmgren Design

Tests the **two** null hypotheses for the positive and the unselected population

$$H_0^- : p^- = p_0 \quad \& \quad H_0^+ : p^+ = p_0$$

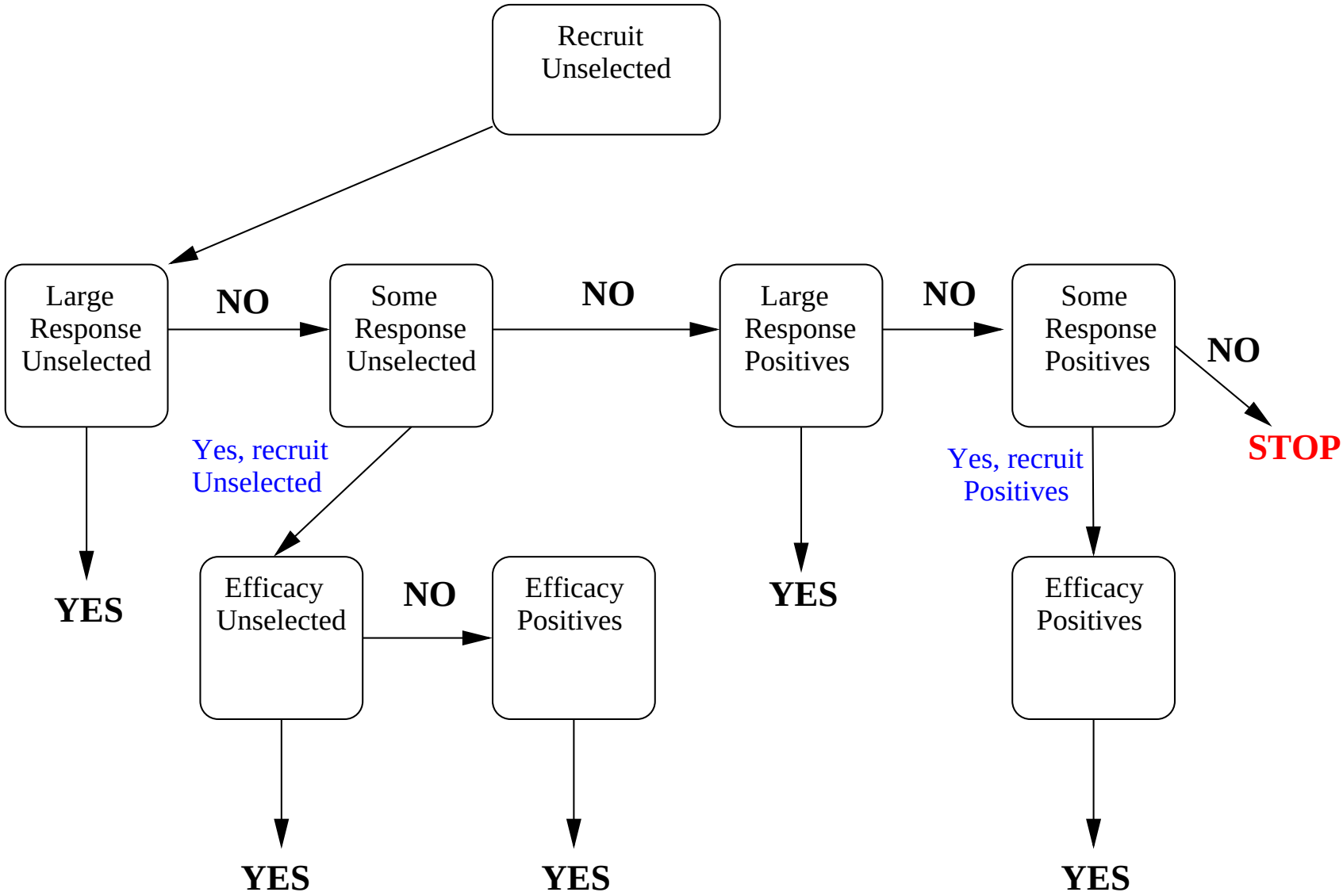
- If you reject H_0^-
  - Conclude efficacy in unselected population
- If you reject H_0^+
  - Conclude efficacy in biomarker positive population

They assume that the response $p^+ > p^-$

Our design parameters

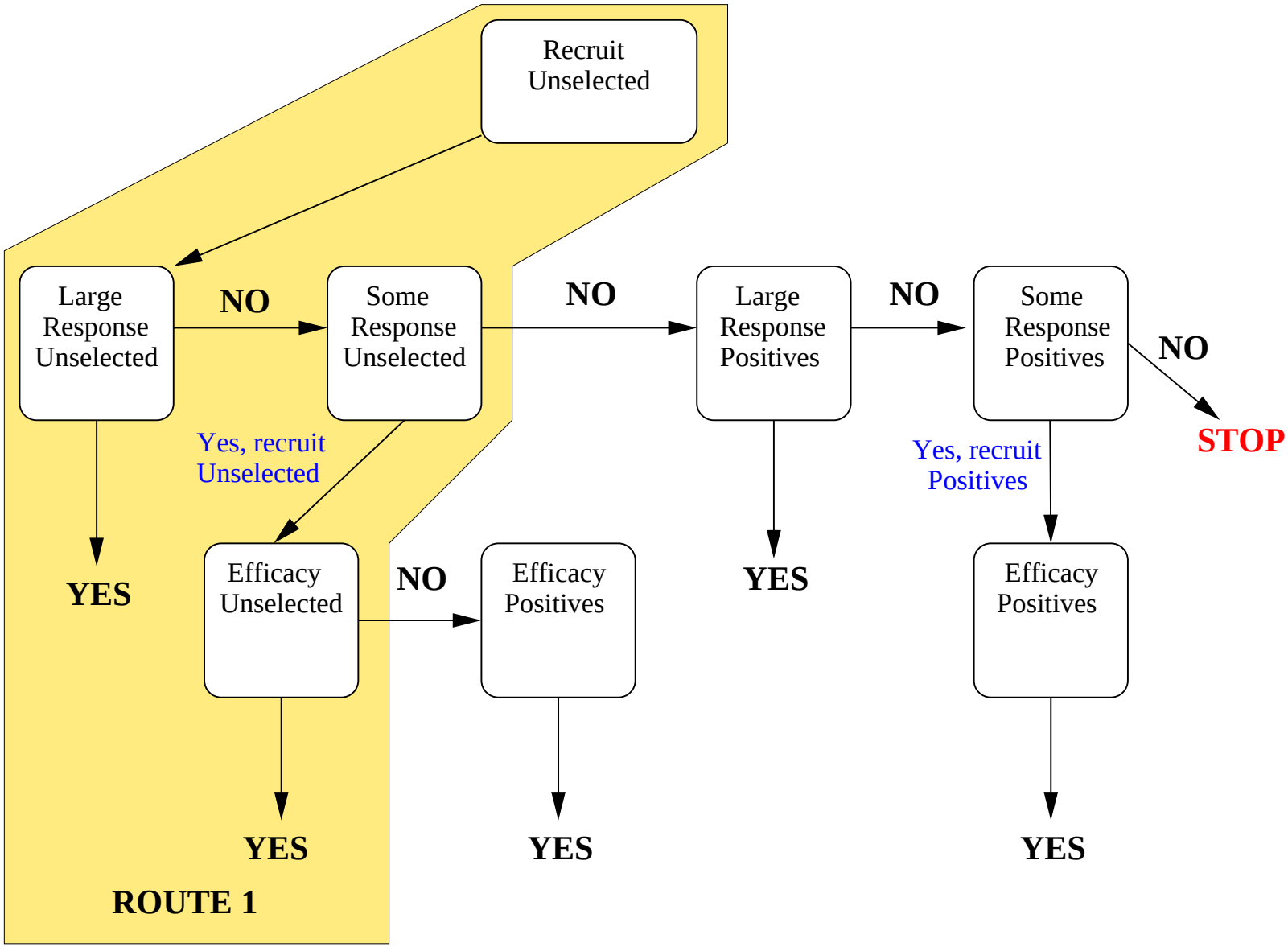
- $p_0 = 0.05$ (under null biomarker is not prognostic)
- 5% significance and 80% power

J&H schematic



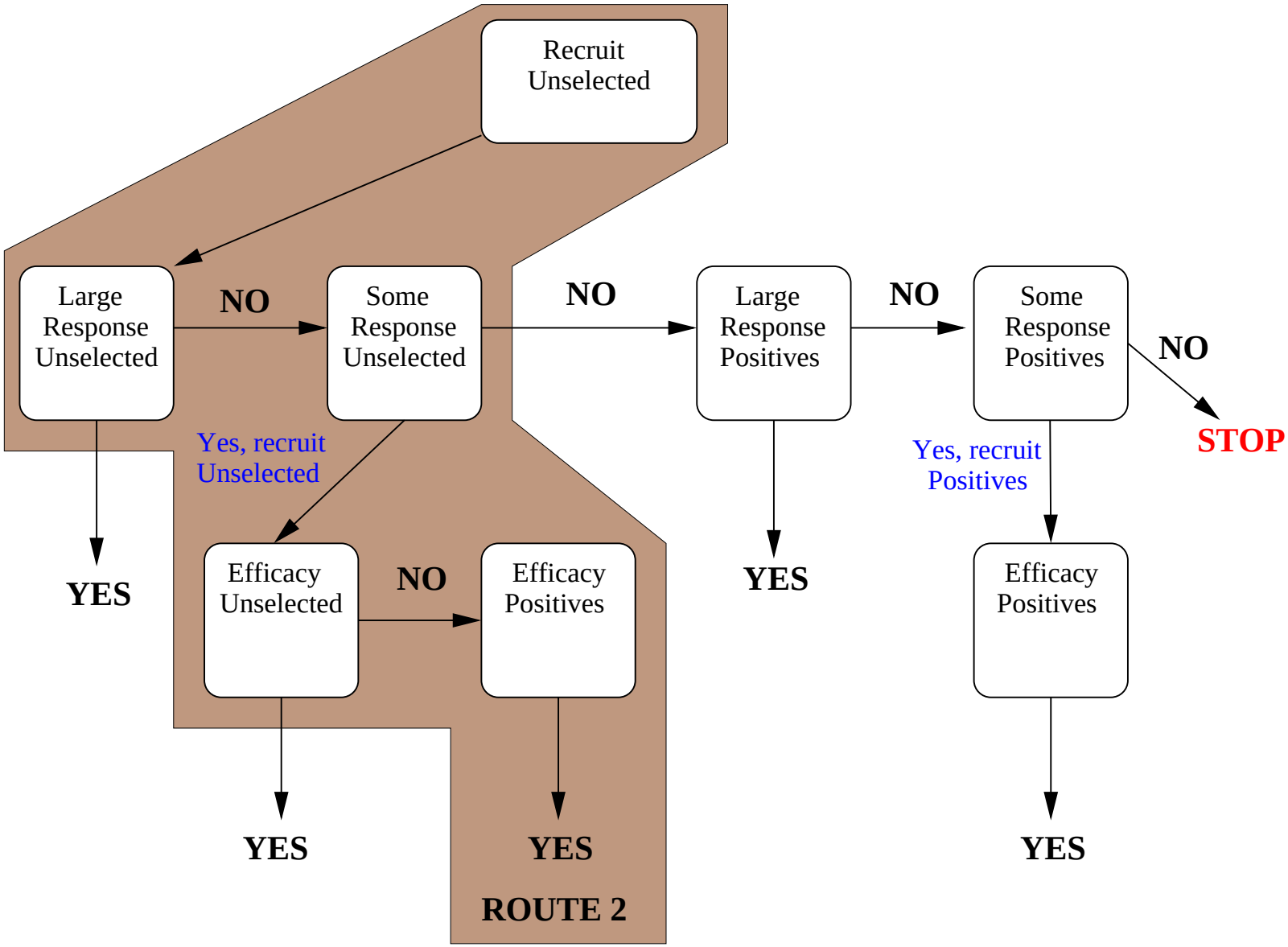
Route 1 - conclusions for the unselected population

Positives not looked at



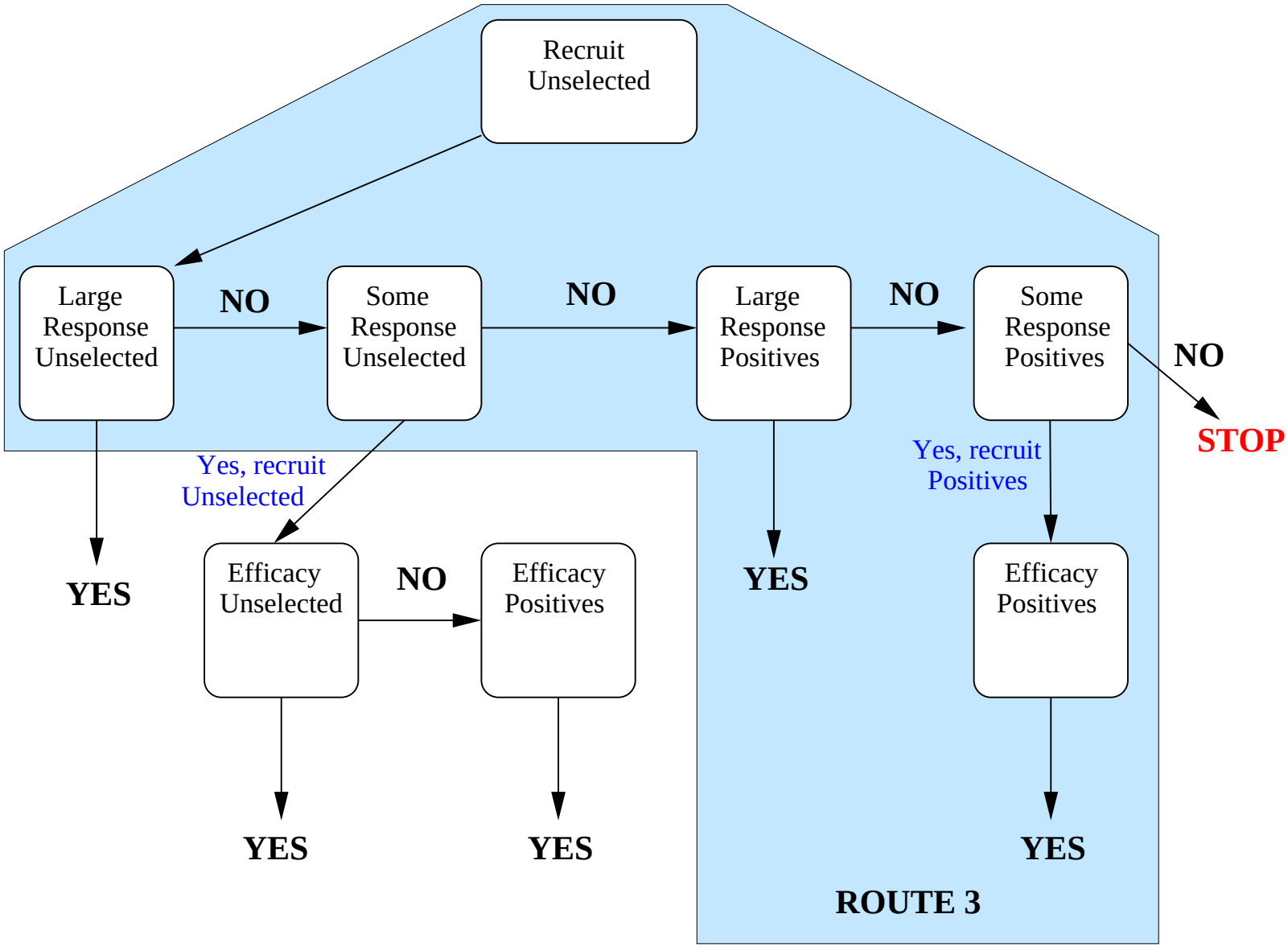
Route 2 - conclusions in positive population

via unselected recruitment



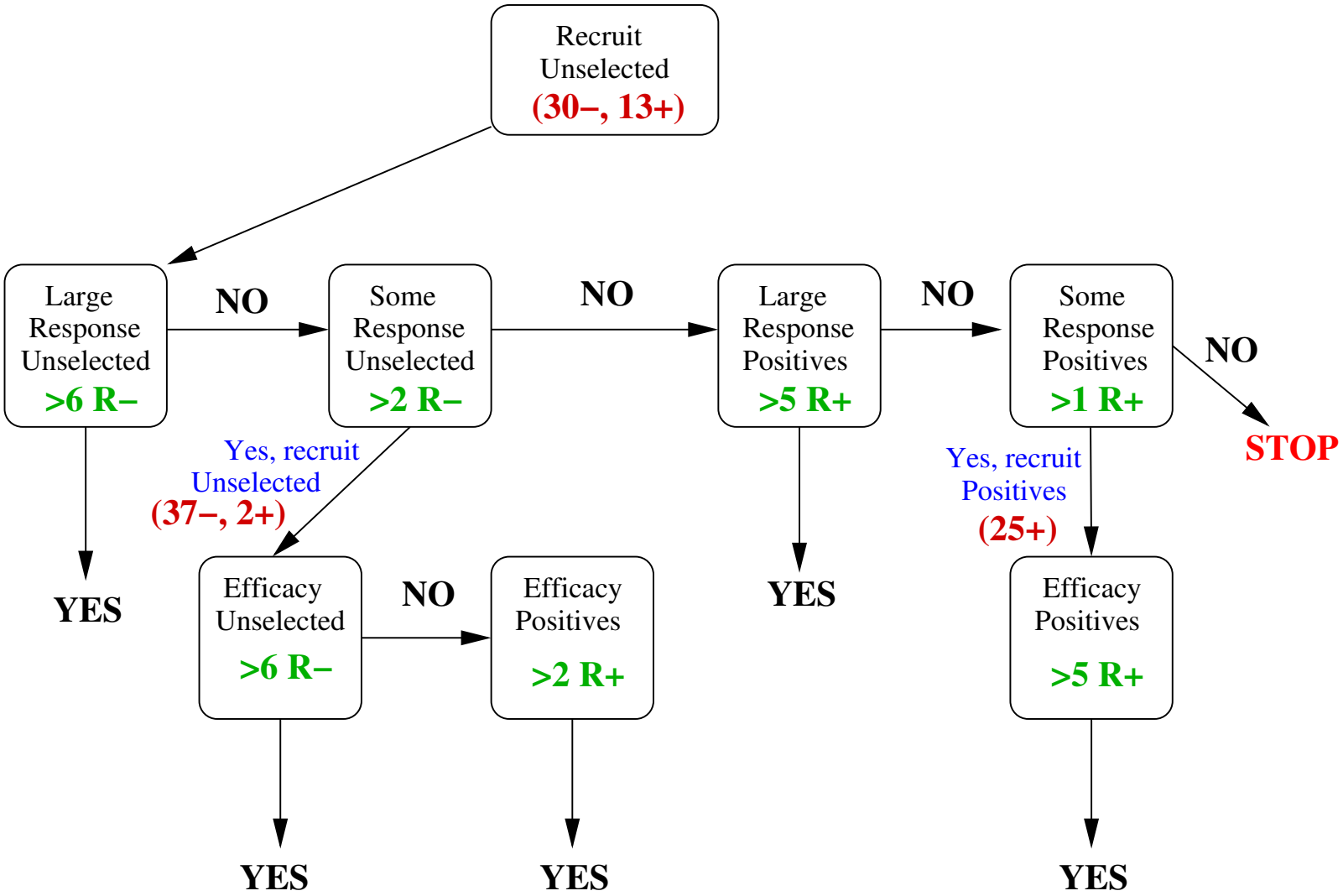
Route 3 - conclusions in positive population

via enrichment sampling



An actual design

$H_0^- : p = 0.05 \quad H_0^+ : p = 0.05$



Shorthand for the design

We characterise all the design parameters as

- Stage 1
  - (3 2)/(30 13)
- Stage 2
  - (6/38) OR (7 3)/(67 15)

There are 10 numbers to find

- with 5 choices for each gives 10 million designs.
- We have fast programs to search a huge design space (Colin) to find best design searching 10 billion designs

Shorthand for the design

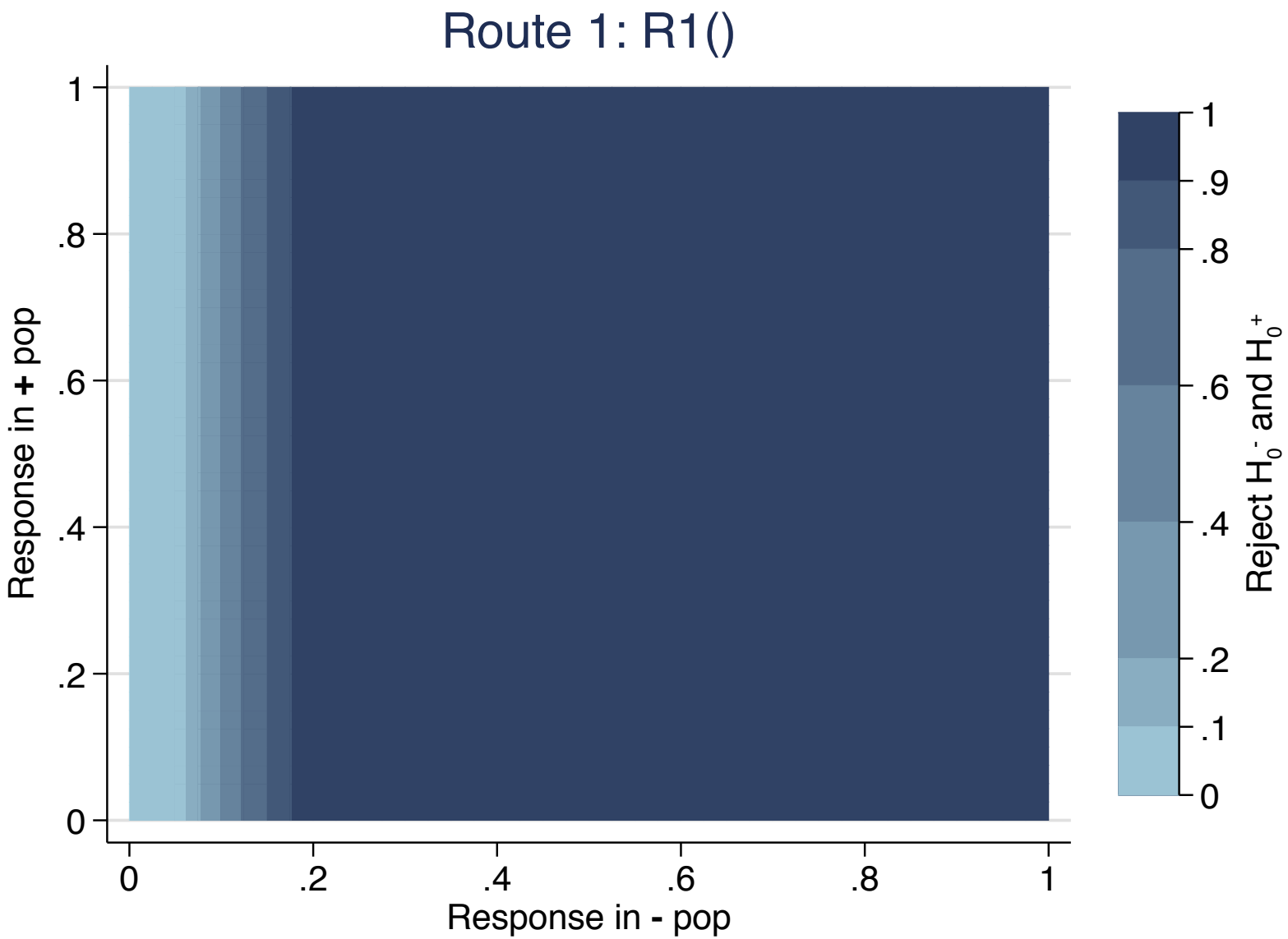
We characterise all the design parameters as

- Stage 1
  - (3 2)/(30 13)
- Stage 2
  - (6/38) OR (7 3)/(67 15)

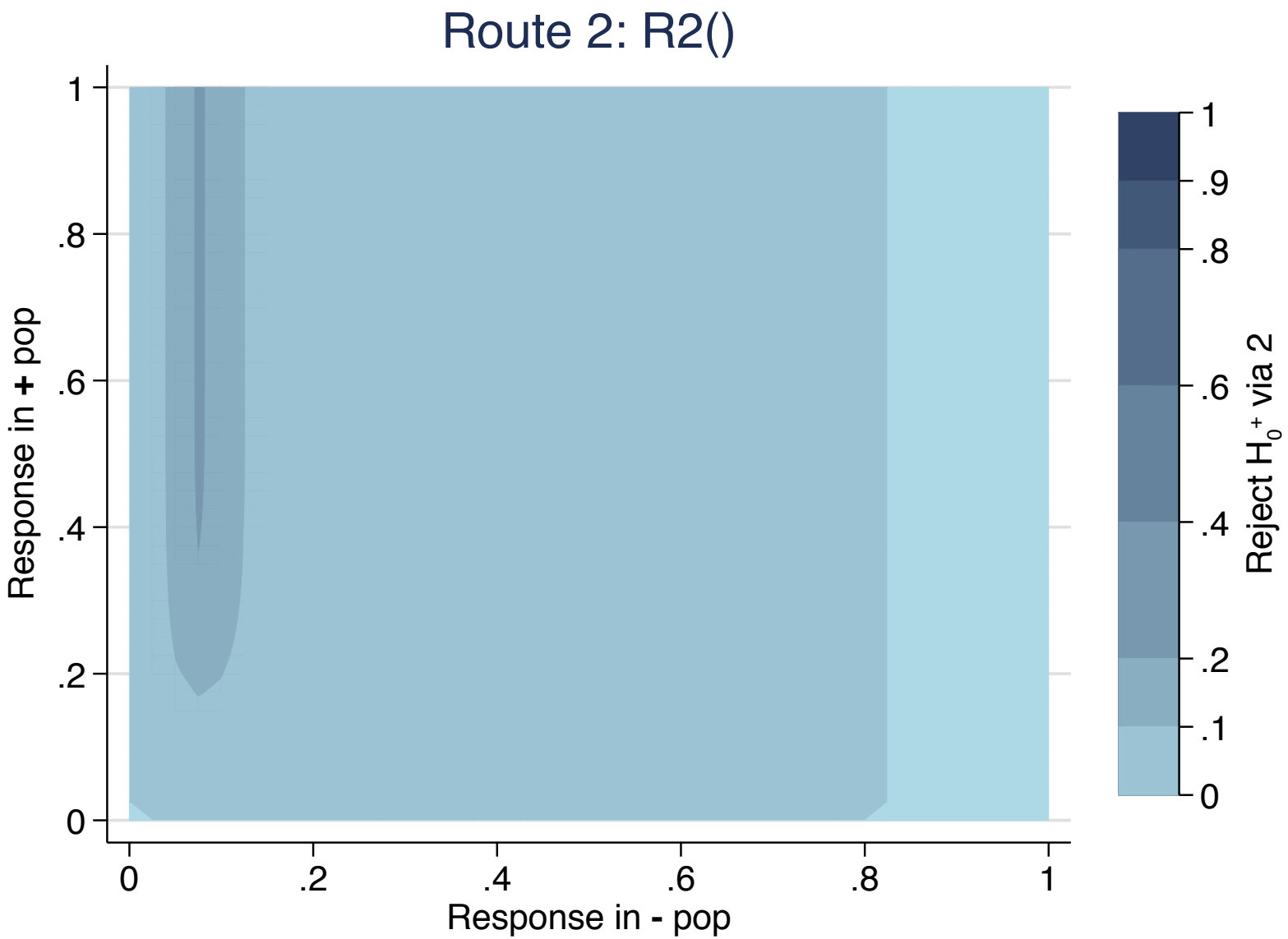
There are 10 numbers to find

- with 5 choices for each gives 10 million designs.
- We have fast programs to search a huge design space (Colin) to find best design searching 10 billion designs

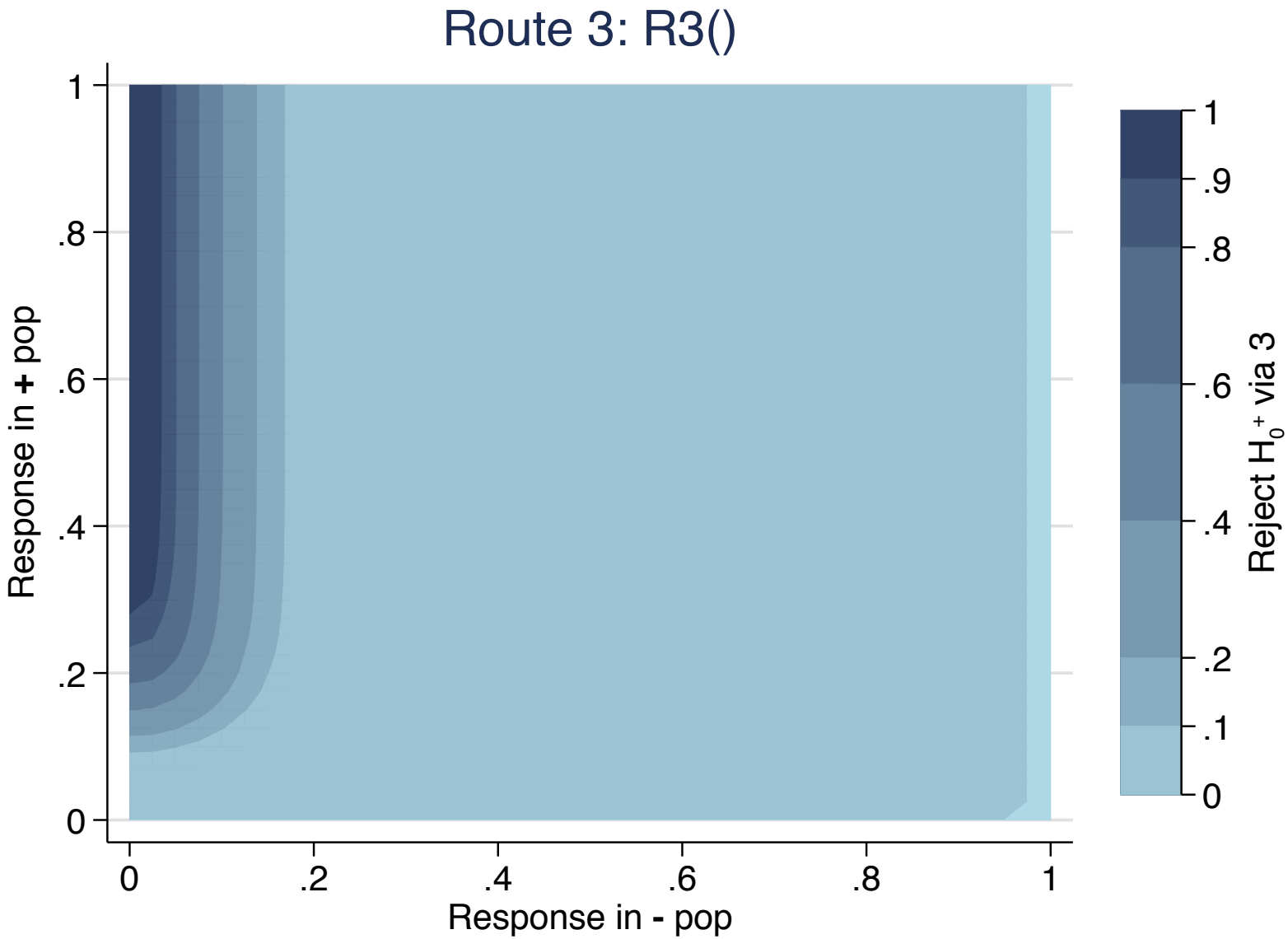
The rejection probabilities for Route 1



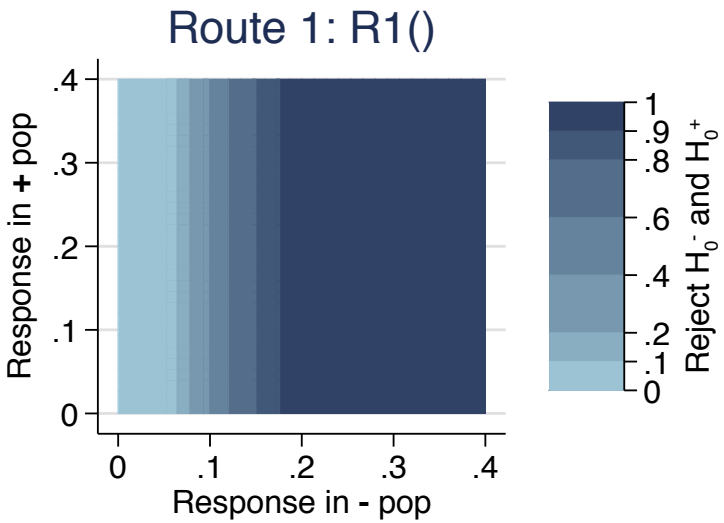
The rejection probabilities for Route 2

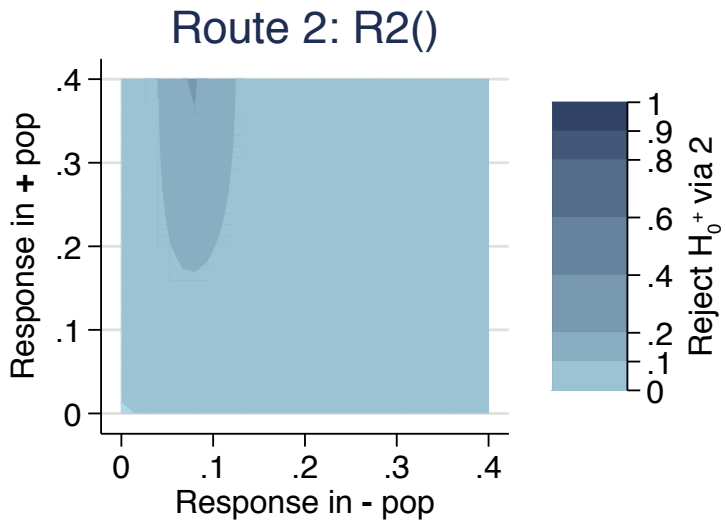


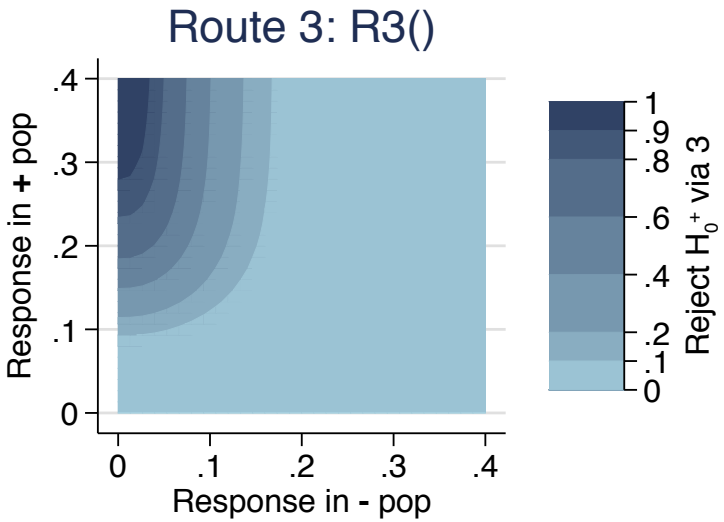
The rejection probabilities for Route 3 (enriched)

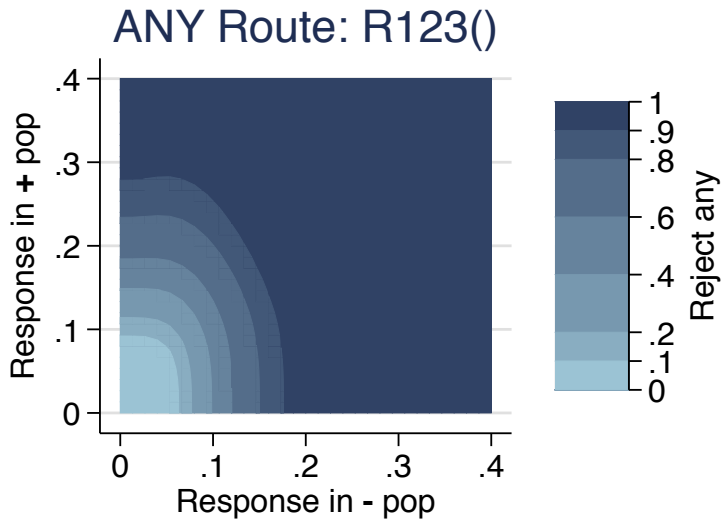


The rejection probabilities









What do the R()s mean?

- $R1(p^-, p^+)$ is the probability of rejecting both nulls via route 1
- $R2(p^-, p^+)$ is the probability of rejecting H_0^+ via route 2
- $R3(p^-, p^+)$ is the probability of rejecting H_0^+ via route 3 (enrichment)

- $R23() = R2() + R3()$
- $R123() = R1() + R2() + R3()$

Power

For this design interest in having enough power when

$$p^- = 0.15 \quad \text{and/or} \quad p^+ = 0.25$$

What do the R()s mean?

- $R1(p^-, p^+)$ is the probability of rejecting both nulls via route 1
- $R2(p^-, p^+)$ is the probability of rejecting H_0^+ via route 2
- $R3(p^-, p^+)$ is the probability of rejecting H_0^+ via route 3 (enrichment)

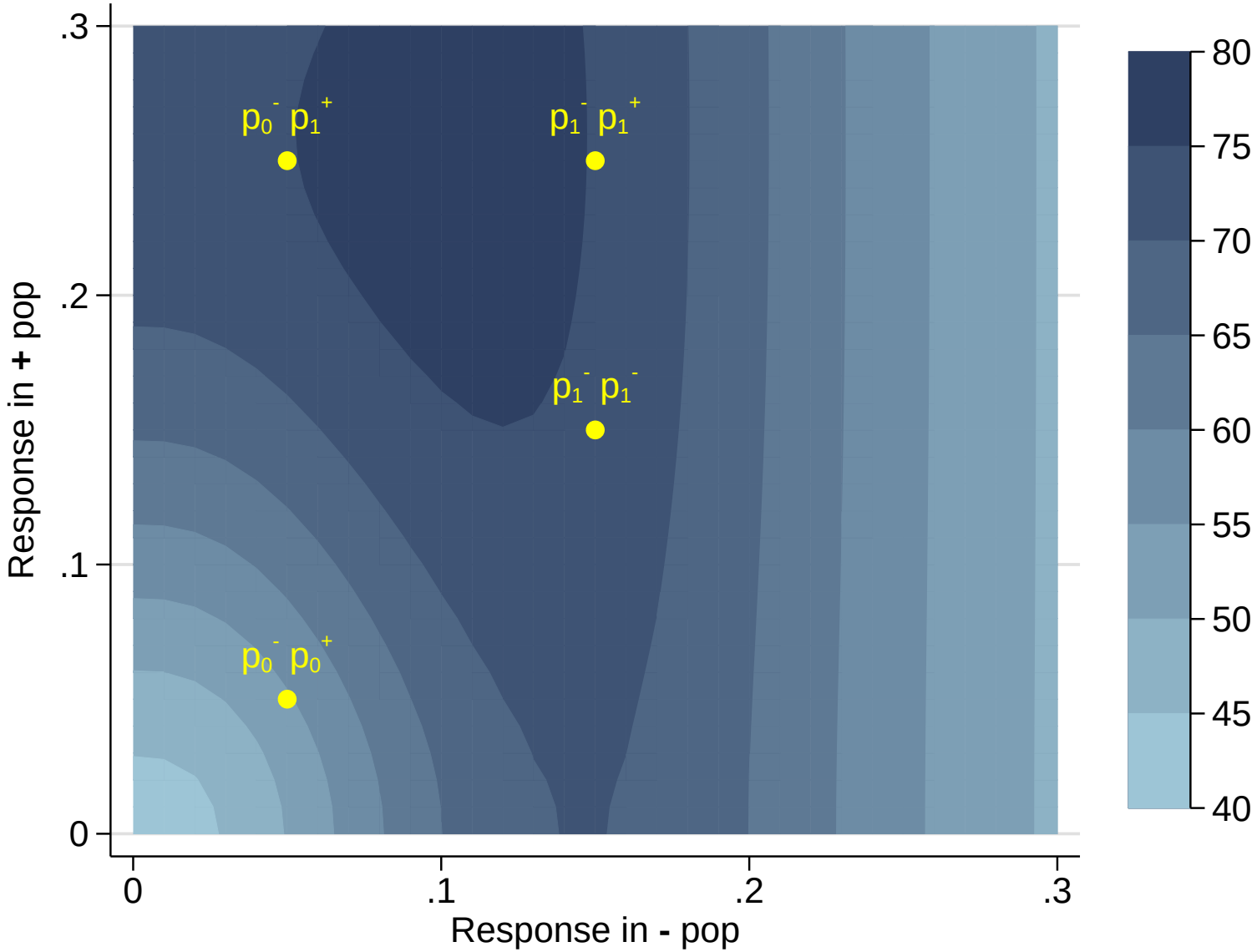
- $R23() = R2() + R3()$
- $R123() = R1() + R2() + R3()$

Power

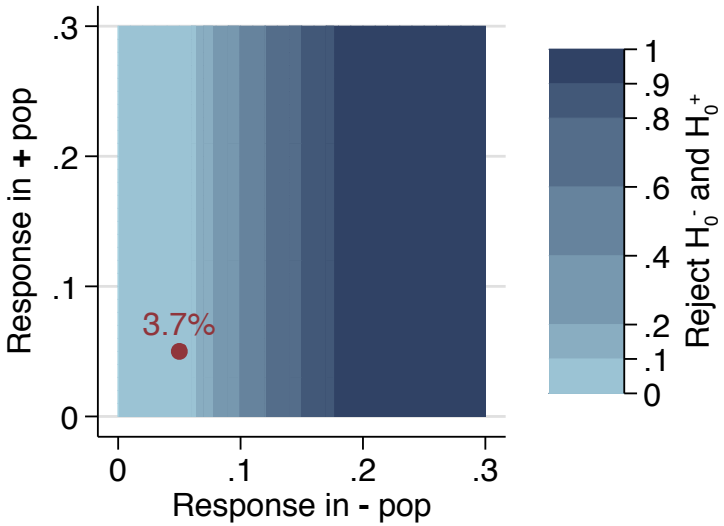
For this design interest in having enough power when

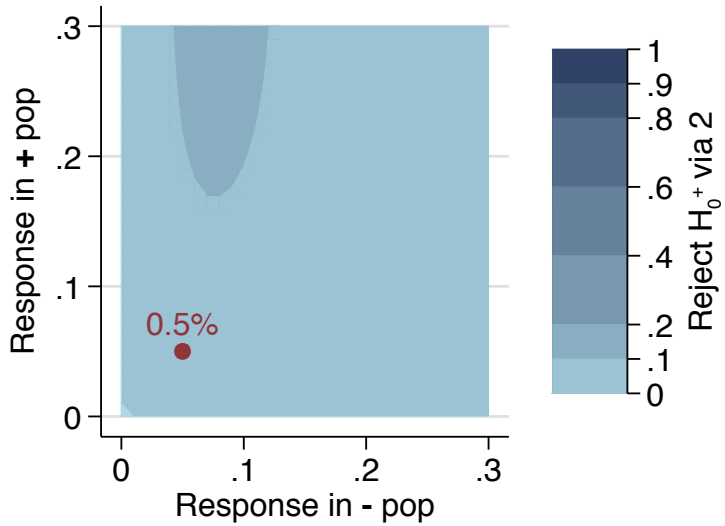
$$p^- = 0.15 \quad \text{and/or} \quad p^+ = 0.25$$

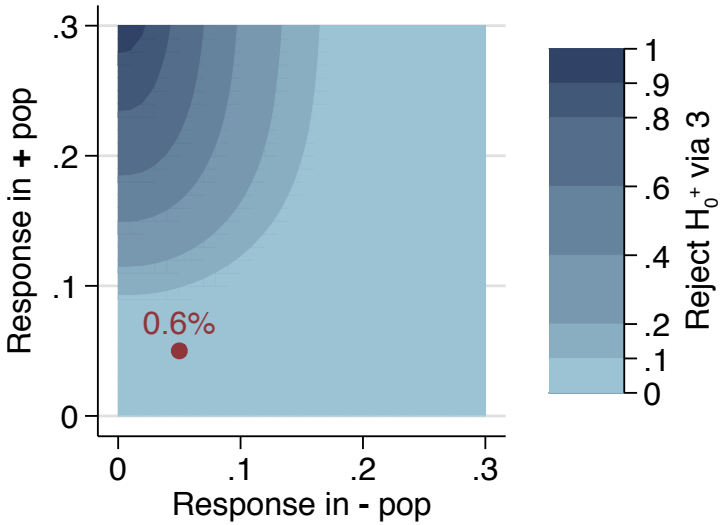
Expected Sample Size

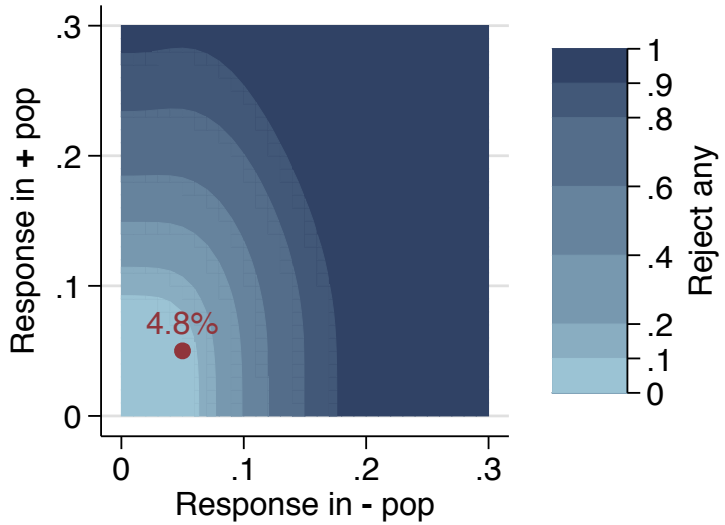


Rejection probabilities at H_0^- & H_0^+









Type 1 error : False positives

Our definition was total error was controlled

$$R123(0.05) \leq 5\% \quad \text{significance}$$

Others could be

Control each error

- $R1(0.05, 0.05) \leq 2.5\%$ and $R23(0.05, 0.05) \leq 2.5\%$
- $R1(0.05, 0.05) \leq 5\%$ and $R23(0.05, 0.05) \leq 5\%$

The first is stronger control than the total control and the latter is weaker. Either are possible.

Type 1 error : False positives

Our definition was total error was controlled

$$R123(0.05) \leq 5\% \quad \text{significance}$$

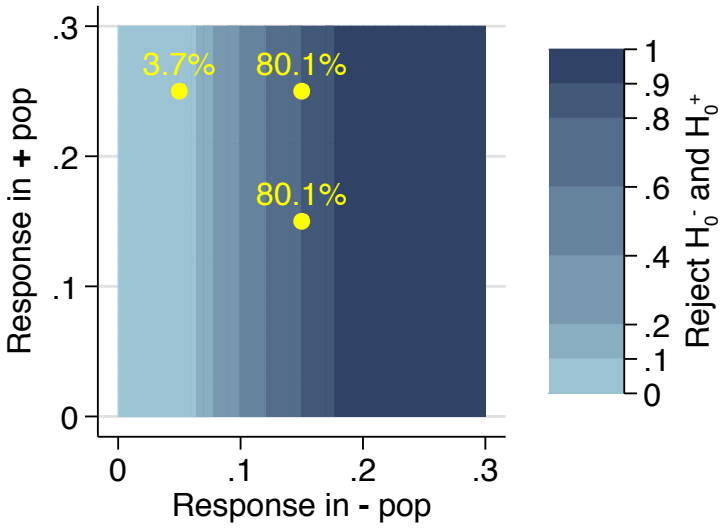
Others could be

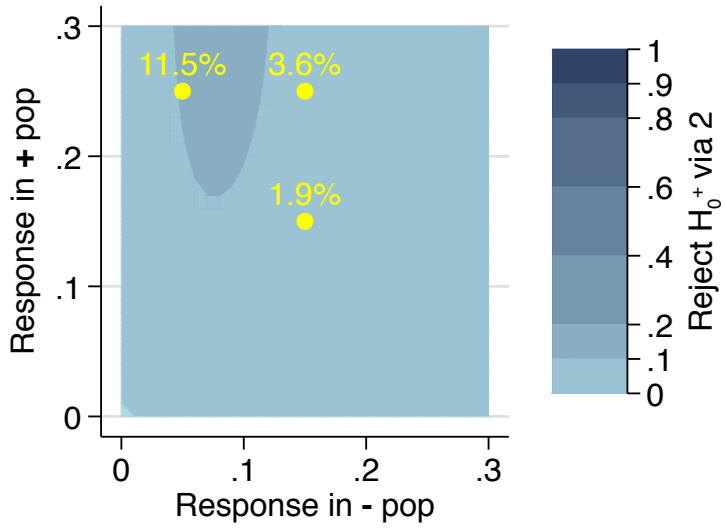
Control each error

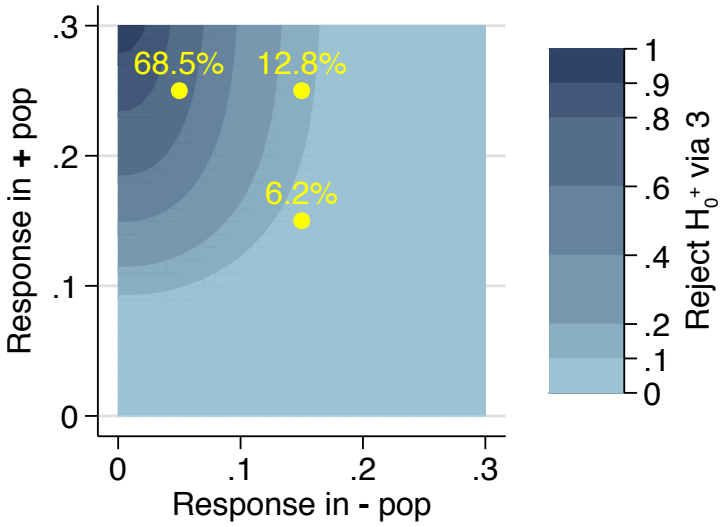
- $R1(0.05, 0.05) \leq 2.5\%$ and $R23(0.05, 0.05) \leq 2.5\%$
- $R1(0.05, 0.05) \leq 5\%$ and $R23(0.05, 0.05) \leq 5\%$

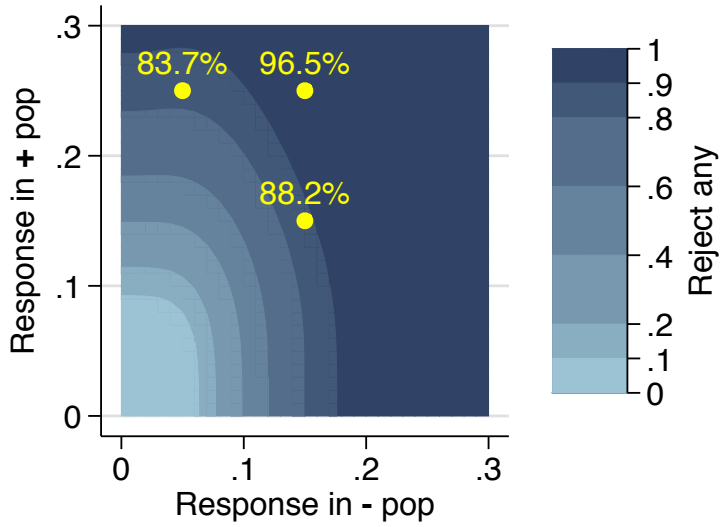
The first is stronger control than the total control and the latter is weaker: Either are possible.

Rejection probabilities at some alternatives









Power

One of our definitions of power was

$$\text{Min}(R1(\text{any}, 0.15), R23(0.25, 0.05)) \geq 80\% \text{ power}$$

The error probabilities can be summarised in a table

	$R1()$	R23()
Null (p_0^-, p_0^+)	$\sum \leq 5\%$	
Unselected (p_1^-, p_1^-)	$\geq 80\%$	
Positive only (p_0^-, p_1^+)		$\geq 80\%$

Therefore our Familywise Error Rate is only weakly controlled

	$R1()$	R23()	R123()
Null (p_0^-, p_0^+)	3.7%	1.1%	4.8%
Unselected (p_1^-, p_1^-)	80.1%	8.1%	
Positive only (p_0^-, p_1^+)	3.7%	80%	

There is 8.1% chance for a wrong positive in unselected

Power

One of our definitions of power was

$$\text{Min}(R1(\text{any}, 0.15), R23(0.25, 0.05)) \geq 80\% \text{ power}$$

The error probabilities can be summarised in a table

	$R1()$	R23()
Null (p_0^-, p_0^+)	$\sum \leq 5\%$	
Unselected (p_1^-, p_1^-)	$\geq 80\%$	
Positive only (p_0^-, p_1^+)		$\geq 80\%$

Therefore our Familywise Error Rate is only weakly controlled

	$R1()$	R23()	R123()
Null (p_0^-, p_0^+)	3.7%	1.1%	4.8%
Unselected (p_1^-, p_1^-)	80.1%	8.1%	
Positive only (p_0^-, p_1^+)	3.7%	80%	

There is 8.1% chance for a wrong positive in unselected

Other possible error controls

The main one is Familywise Error Rate being strongly controlled

	$R1()$	R23()
Null (p_0^-, p_0^+)	$\sum \leq 5\%$	
Unselected (p_1^-, p_1^-)	$\geq 80\%$	$\leq 5\%$
Positive only (p_0^-, p_1^+)	$\leq 5\%$	$\geq 80\%$

With stronger false positive control we get

	$R1()$	R23()	R123()
Null (p_0^-, p_0^+)	$\leq 2.5\%$	$\leq 2.5\%$	$\leq 5\%$
Unselected (p_1^-, p_1^-)	$\geq 80\%$	$\leq 5\%$	
Positive only (p_0^-, p_1^+)	$\leq 5\%$	$\geq 80\%$	

Other possible error controls

The main one is Familywise Error Rate being strongly controlled

	$R1()$	R23()
Null (p_0^-, p_0^+)	$\sum \leq 5\%$	
Unselected (p_1^-, p_1^-)	$\geq 80\%$	$\leq 5\%$
Positive only (p_0^-, p_1^+)	$\leq 5\%$	$\geq 80\%$

With stronger false positive control we get

	$R1()$	R23()	R123()
Null (p_0^-, p_0^+)	$\leq 2.5\%$	$\leq 2.5\%$	$\leq 5\%$
Unselected (p_1^-, p_1^-)	$\geq 80\%$	$\leq 5\%$	
Positive only (p_0^-, p_1^+)	$\leq 5\%$	$\geq 80\%$	

Conclusions

- We belief you want to control the wrong positive error
- We optimised with respect to the expected sample size under the global null
- We have software that used massive parallelisation
- Future — want to understand whether there is a role for single-arm trials in biomarker trials

Conclusions

- We belief you want to control the wrong positive error
- We optimised with respect to the expected sample size under the global null
- We have software that used massive parallelisation
- Future — want to understand whether there is a role for single-arm trials in biomarker trials

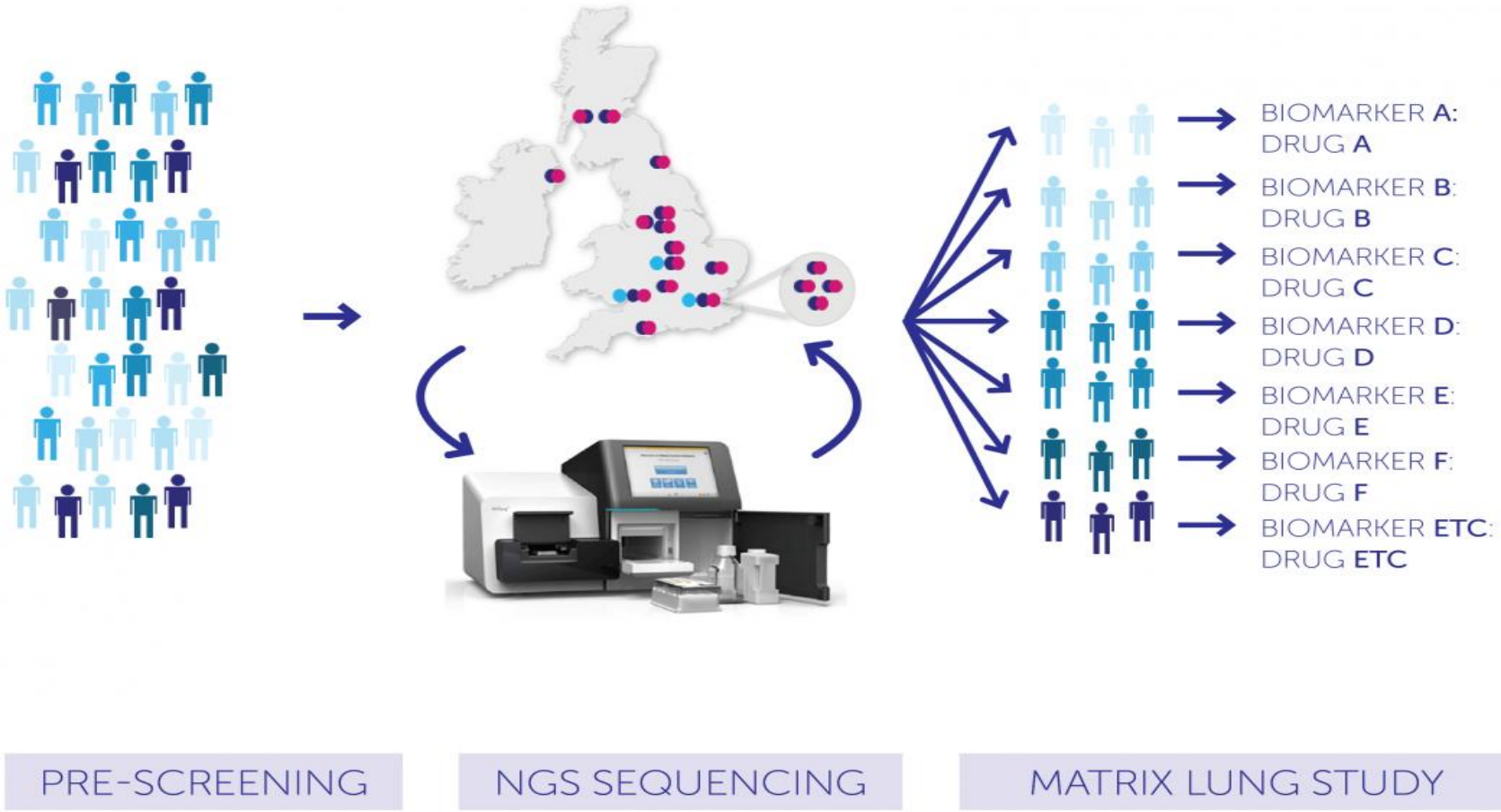
Multiple experimental treatments

- If there are several experimental treatments available for testing, then there are substantial advantages of including several arms in a single 'umbrella' trial.
- A shared control group means more statistical efficiency: test more treatments with the same number of centres.
- Administratively and logistically easier compared to separate trials.
- For targeted treatments: more enrolled patients will receive a treatment targeted at their biomarker profile.
- However, also these types of trials are also more complicated.

Design 1: Parallel trials

- One type of platform trial consists of a set of parallel trials.
- A patient is allocated to a trial on the basis of their biomarker profile.
- A couple of UK examples:
  - National lung matrix trial
  - FOCUS 4
- Both of these also use adaptive design approaches to stop sub-trials where the treatment is not showing sufficient signs of efficacy.

Design 1: Parallel trials



Design 2: Bayesian adaptive randomisation

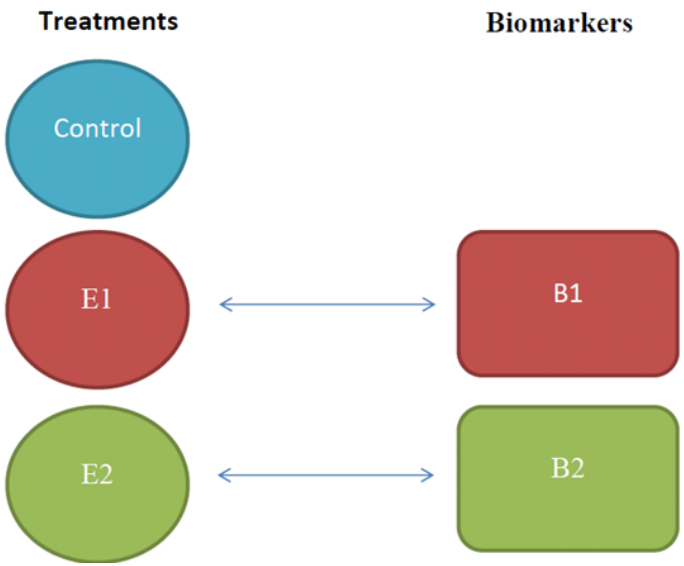
- A second type of umbrella trial does not make assumptions of links between biomarkers and treatments.
    - Example: BATTLE, ISPY2
- Both of these use Bayesian adaptive randomisation (BAR) to change the randomisation probability:
- A patient is more likely to receive treatments that have previously worked well on patients with similar biomarker profiles.

Design 3: Linked BAR design

- When the links between treatment and biomarker are plausible but unsure, neither design seems completely appropriate.
- Intermediate choice: linked-BAR design[1].
- Combines initial stage of parallel-trials design then uses BAR to update allocation in case alternative links are present.

[1] Wason J, Abraham J, Baird R, Gournaris J, Vallier A, Brenton J, Earl H, Mander A. (2015) A Bayesian adaptive design for biomarker trials with linked treatments. British Journal of Cancer 113, 699-705

Multi-arm trial



- Each experimental treatment may 'linked'with one of the biomarkers.
- Treatments thought likely to work well for patients with linked biomarker.
- Not known: treatment may work in a broader set of patients (or in none).
- Several designs available for this scenario

Motivating clinical example - post-adjuvant breast cancer

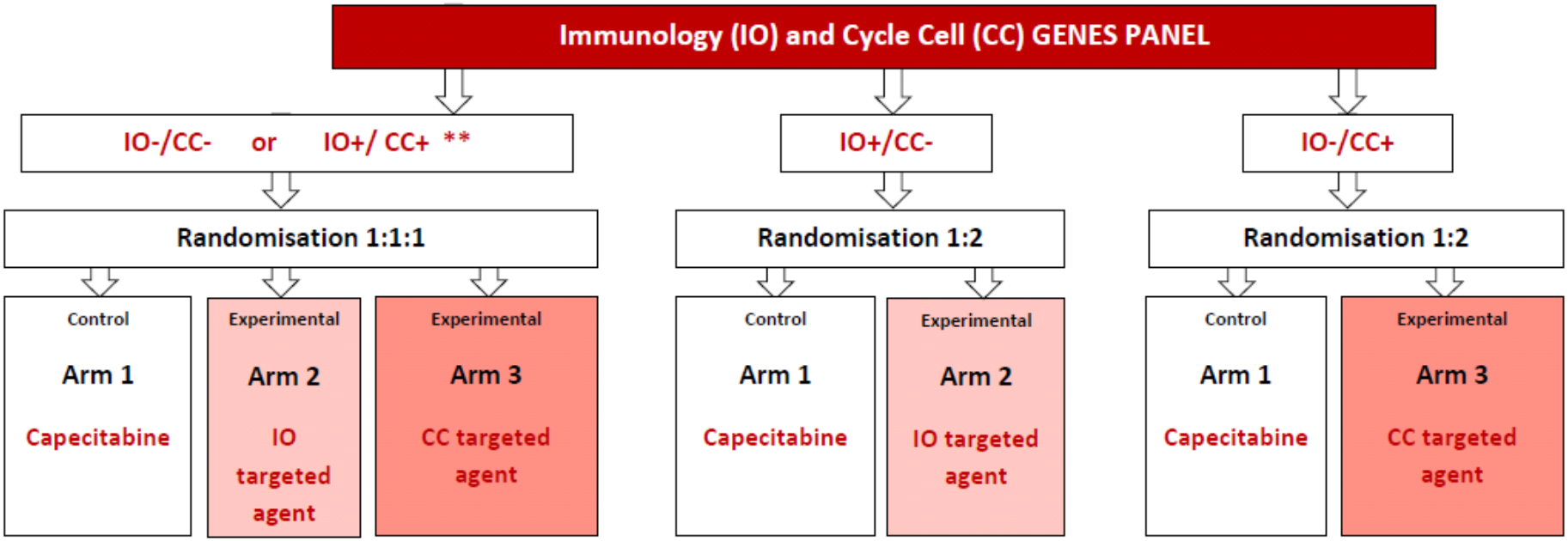
- Japanese trial has recently shown that capecitabine can improve long-term disease-free survival after breast surgery in poor prognosis groups.
- Clinical collaborators in University of Cambridge oncology department wanted phase II design that would test whether targeted agents would offer advantages over capecitabine.
- Patient population is women who have residual circulating tumour DNA after tumour removal operation — data shows this group has poor prognosis.

Motivating clinical example - post-adjuvant breast cancer

- Primary endpoint — log percentage change in circulating tumour DNA level from baseline to six months. Immunology and cycle cell gene panels used as the biomarkers.
- Moderately prevalent biomarkers (30% for each).
- Treatment arms include capecitabine (control) and two targeted agents that would be thought to work in patients who have high levels of the relevant gene panel.

Linked BAR design

- Stage 1: 100 patients recruited and randomised between control and experimental arm linked with a biomarker the patient is positive for. Control arm randomisation is always 1/3.
  - E.g. if patient is positive for biomarker 1, randomised between control and treatment 1 in 1:2 ratio.
  - If patient positive for both biomarkers or neither, randomised 1:1:1 between control and experimental treatments.



Model used for BAR

- Stage 2 (200 patients): at a series of interim analyses, recommended allocation probabilities get updated according to results so far.
- Bayesian linear model fitted at each interim. Model contains intercept, marginal effects of each experimental treatment (β), marginal effect of each biomarker (γ), and interactions between biomarkers and treatment (δ).

$$\log\left(\frac{y_{i1}}{y_{i0}}\right) = \mu + \beta_{T(i)} + \sum_{j=1}^{2} \gamma_j x_{ij} + \sum_{j=1}^{2} \delta_{T(i)j} x_{ij} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where y_{i0} and y_{i1} are ctDNA measurements at baseline and six months respectively, $T(i)$ is allocated treatment of patient i, x_{ij} is 1 if patient i is positive for biomarker j.

Linked BAR model

- Model uses normal-inverse gamma form for conjugacy.
- All parameters except δ_{11} and δ_{22} have non-informative priors.
- δ_{11} and δ_{22} have moderately informative priors chosen to continue favouring allocation of patients to linked treatments (until there is sufficient evidence that they are not working).
- Model gives posterior probability of each experimental treatment being superior to control for each possible biomarker profile.
- These posterior probabilities are then transformed into allocation probabilities for future patients (see Wason et al. for more details).

Final Analysis

After all patients have been assessed, (frequentist) linear regression is fitted with same parameters as previously.

$$\log\left(\frac{y_{i1}}{y_{i0}}\right) = \mu + \beta_{T(i)} + \sum_{j=1}^{2} \gamma_j x_{ij} + \sum_{j=1}^{2} \delta_{T(i)j} x_{ij} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

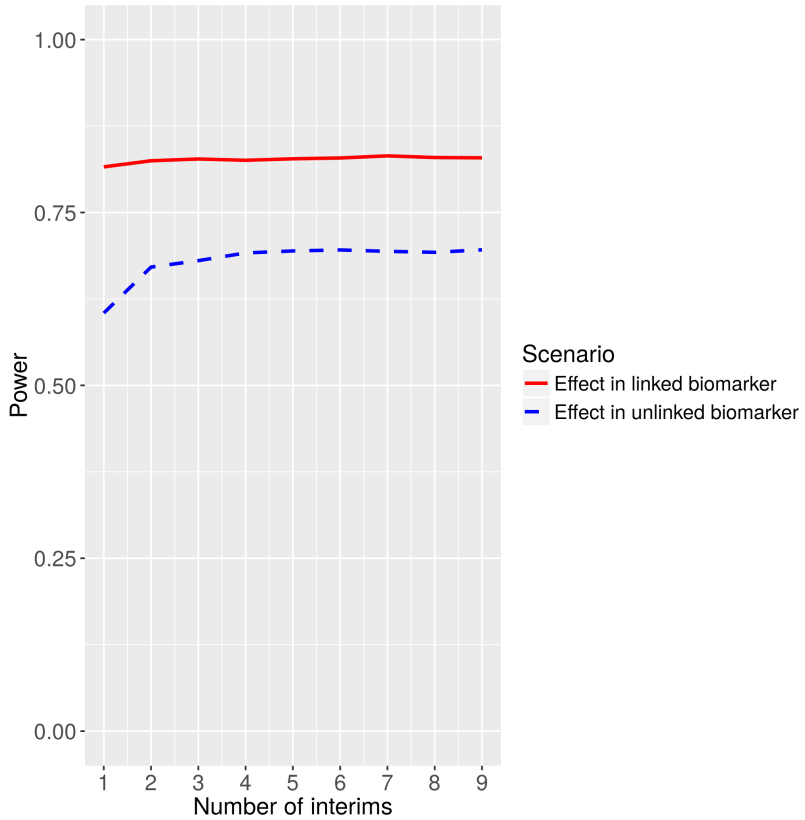
Effect of each experimental treatment can be tested in each biomarker group (and biomarker negative group). E.g. effect of experimental treatment 1 can be tested in biomarker 2 positive patients by testing:

$$H_0^{12} : \beta_1 + \delta_{12} \geq 0$$

Each hypothesis tested at one-sided 5% error rate

How many interim anlayses in stage 2

- Scenario 1 (red) — experimental treatment works in linked biomarker group (standardised effect size 0.65).
- Scenario 2 (blue) — experimental treatment works in non-linked biomarker group.
- Number of interims has low impact on scenario 1 power.
- Increases scenario 2 power, but only up to 4.



Comparison of designs (50000 replicates)

Scenario	Parallel trials power	BAR power	Linked-BAR power
Trt 1 works in all patients	0.949	0.977	0.976
Trt 1 works in biomarker 1 positive patients	0.831	0.798	0.833
Trt 1 works in biomarker 2 positive patients	0.428	0.796	0.699

	Parallel trials	BAR	Linked-BAR
Maximum type I error rate	0.248	0.214	0.212

Conclusions

- Wason et al. shows comparisons for large number of scenarios.
- Generally:
  - When biomarker-treatment links are correct: parallel trials best, linked BAR very close. BAR loses moderate amount of power but still pretty good.
  - When links are incorrect: BAR best, linked BAR loses moderate amount of power; parallel trials low power.

Acknowledgements

All members of the MRC Biostatistics Unit

- Deepak Parashar (Warwick Uni)
- Colin Starr
- Jack Bowden (Bristol Uni)
- Lorenz Wernisch
- James Wason