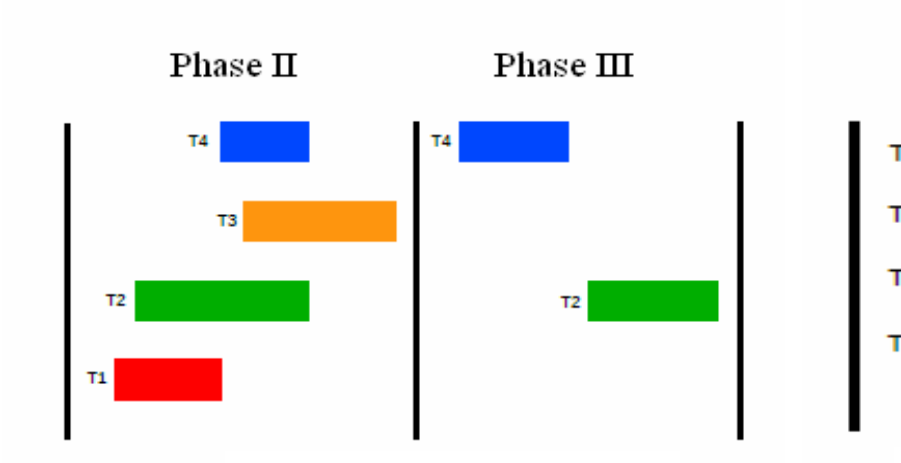
Optimal design of multi-arm multi-stage (MAMS) clinical trials

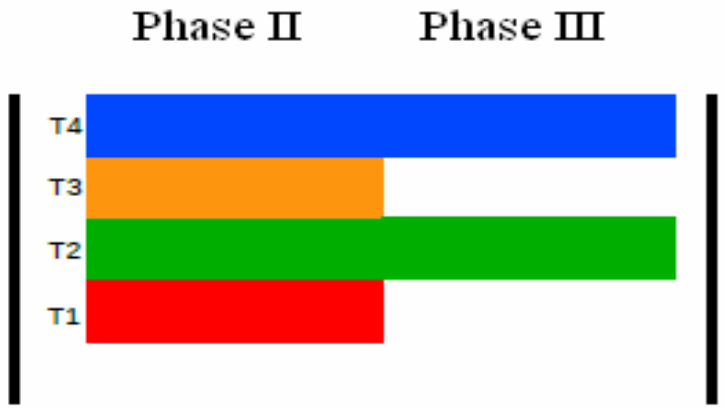
James Wason, Thomas Jaki

Introduction - multi-arm trials

- In some therapeutic areas, there may be several possible agents/treatments awaiting trials.
- The traditional approach is to test each one by one in separate controlled clinical trials.
- An alternative is one trial in which several novel treatments are compared. Advantages:
  - Efficient and cheaper, since a shared control group is used.
  - More treatments can be tested with a limited set of patients.
  - More popular with patients as a greater chance of being allocated to a new treatment.



(a) Traditional



(b) MAMS

MAMS trials

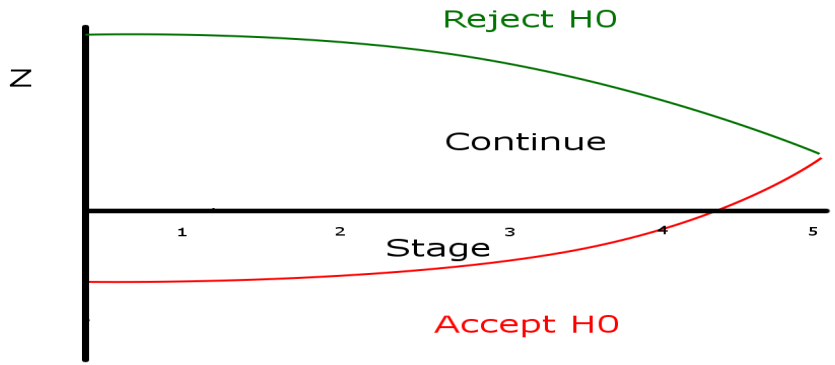
- By introducing interim analyses (multi-stage), further efficiency gained.
- At interim analyses, arms may be dropped if evidence so far suggests they are unlikely to be effective (futility), or if sufficient evidence of effectiveness has been found already (efficacy).
- Multi-arm multi-stage (MAMS) trials are a very broad class of design:
  - Comparisons may be between active treatments and control treatment, or pairwise between all arms.
  - Stopping for efficacy may mean the trial stops, or just the relevant arm (or stopping for efficacy may not be allowed).
  - The endpoint tested may be the same at each interim analysis, or may differ.
- Our motivating example is the TAILOR trial - phase II MAMS trial comparing four active arms to a control arm for the treatment of insulin resistance in HIV-positive individuals. The endpoint is change in insulin resistance as measured by HOMA-IR.

Aims of work

- We wished to explore 'optimal' MAMS designs - i.e. choosing the stopping boundaries to minimise the expected sample size.
- Expected sample size is very important to control as it determines efficiency of drug development process, as well as how many patients are exposed to ineffective treatments.
- This has not been possible up to now, so we have worked on developing methodology to allow exploration of optimal MAMS trials. Our questions of interest were:
  - How do optimal stopping boundaries vary with number of active treatments?
  - How do traditional group-sequential stopping boundaries perform?
  - Should we allocate more patients to the control group; if so, how many more?

Assumptions

- Assumptions from design of TAILOR trial:
  - Change in insulin resistance is used as endpoint at each analysis, and is normally distributed.
  - The variance of the change in insulin resistance outcome is known for each treatment.
  - Comparisons are between each active arm and the control arm.
  - Stopping for futility and efficacy are both allowed. Stopping one arm for efficacy means the whole trial stops.



- Trial design parameterised by $(n, f_1, \ldots, f_J, e_1, \ldots, e_J)$.

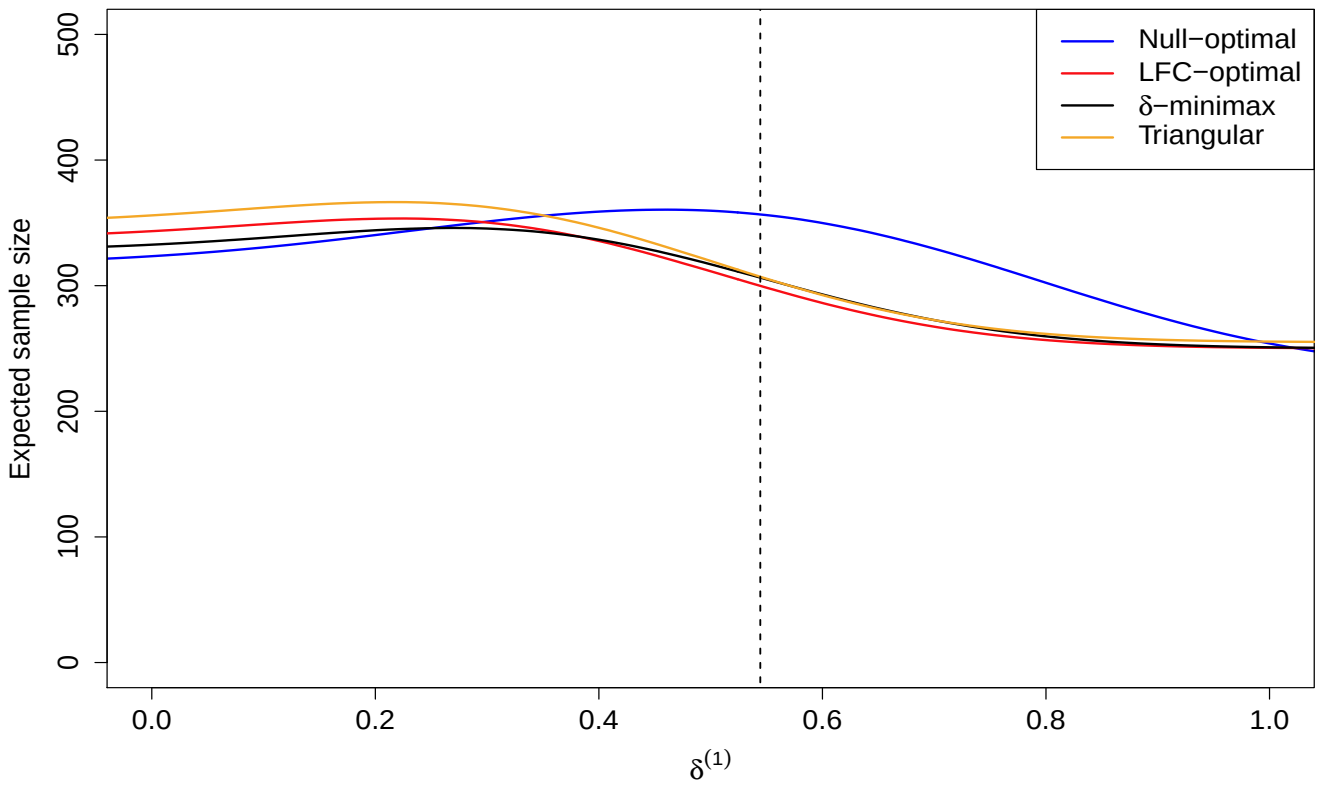
Optimal designs

- An optimal MAMS design minimises expected sample size at some vector of treatment effects with control over type-I error rate and power.
- Infinite number of optimality criteria, so we restrict ourselves to three:
  - Global-null-optimal design: optimal when all treatments are ineffective.
  - LFC-optimal design: optimal when one treatment is effective and others are not.
  - δ-minimax design - optimal when all but one treatment is ineffective and treatment effect of other is picked to give the highest expected sample size.
- All are generalisations of optimal designs from the one active treatment case.
- Finding optimal MAMS designs is difficult and requires computationally intensive techniques (see submitted paper for more details).

Results

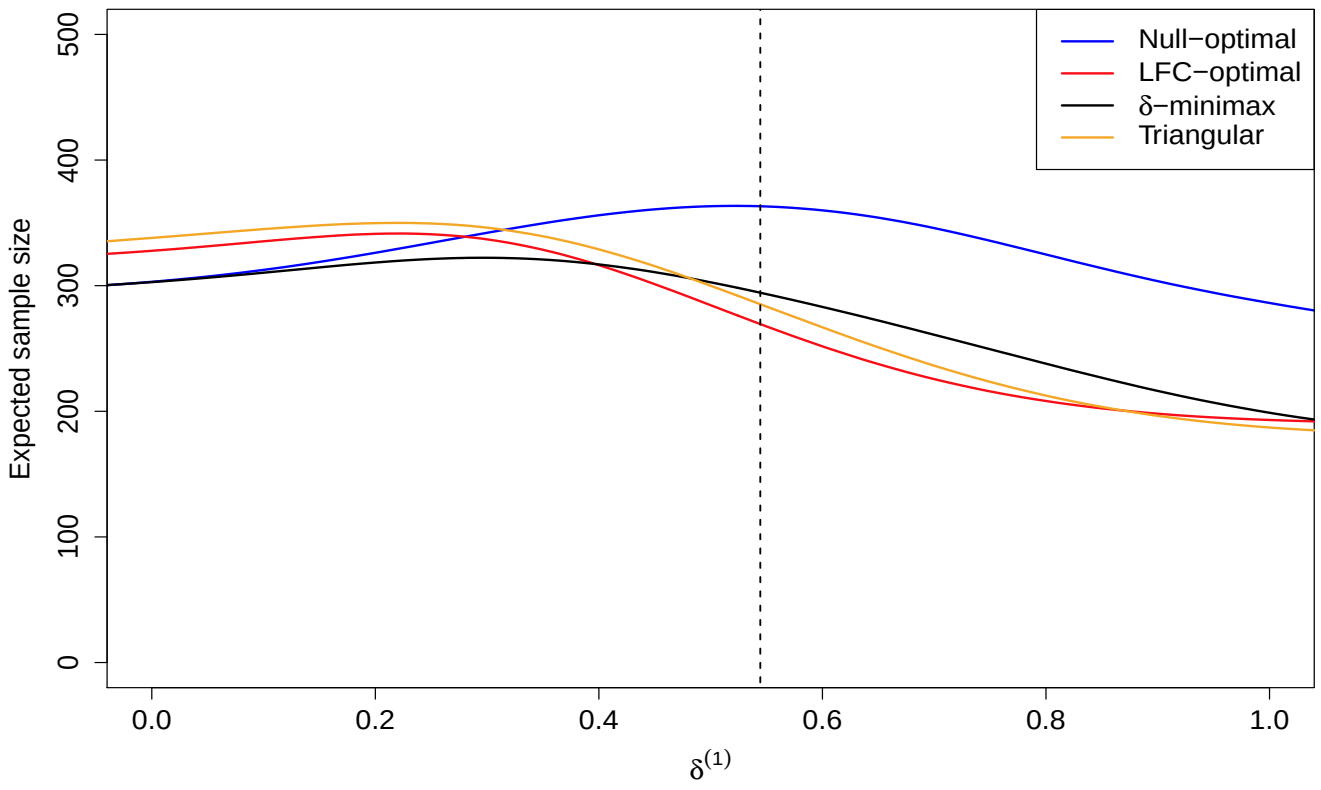
- Firstly we compare designs in terms of expected sample size.
- Four designs examined, the three optimal designs and the triangular test (for one active treatment, triangular test properties rival that of optimal designs in terms of expected sample size).
- Firstly we compare designs when all but one of the active treatments is ineffective, and the treatment effect of the other varies:

Expected sample size (1) (number of active treatments = 4, number of stages = 2)



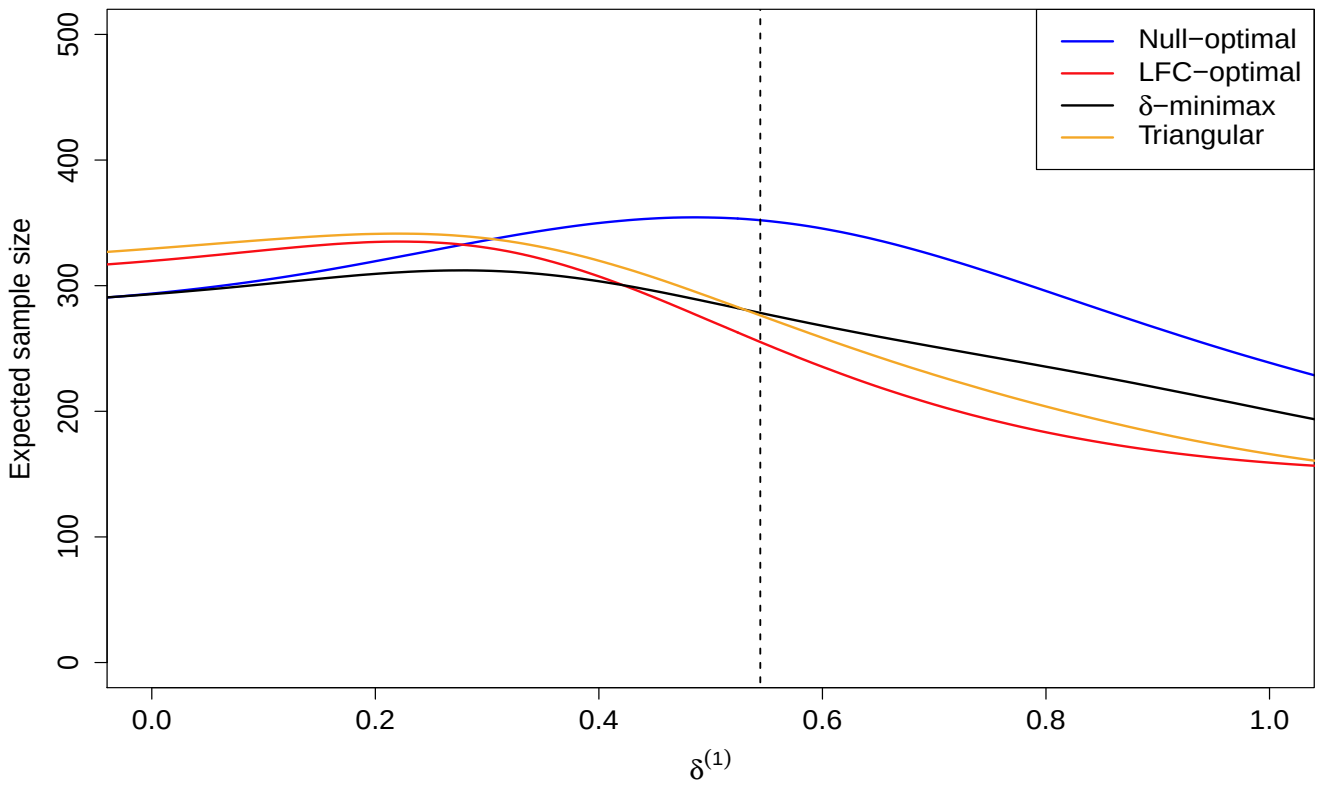
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (1) (number of active treatments = 4, number of stages = 3)



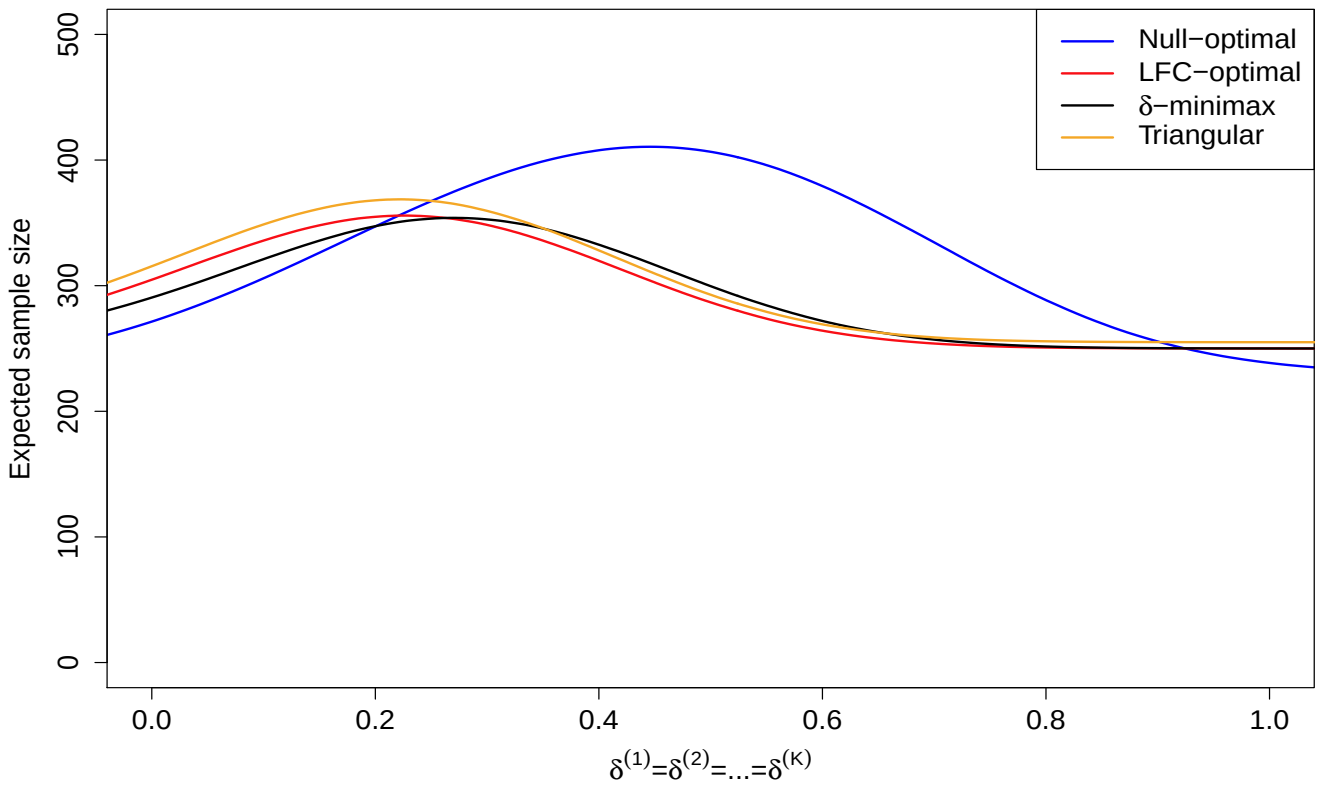
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (1) (number of active treatments = 4, number of stages = 4)



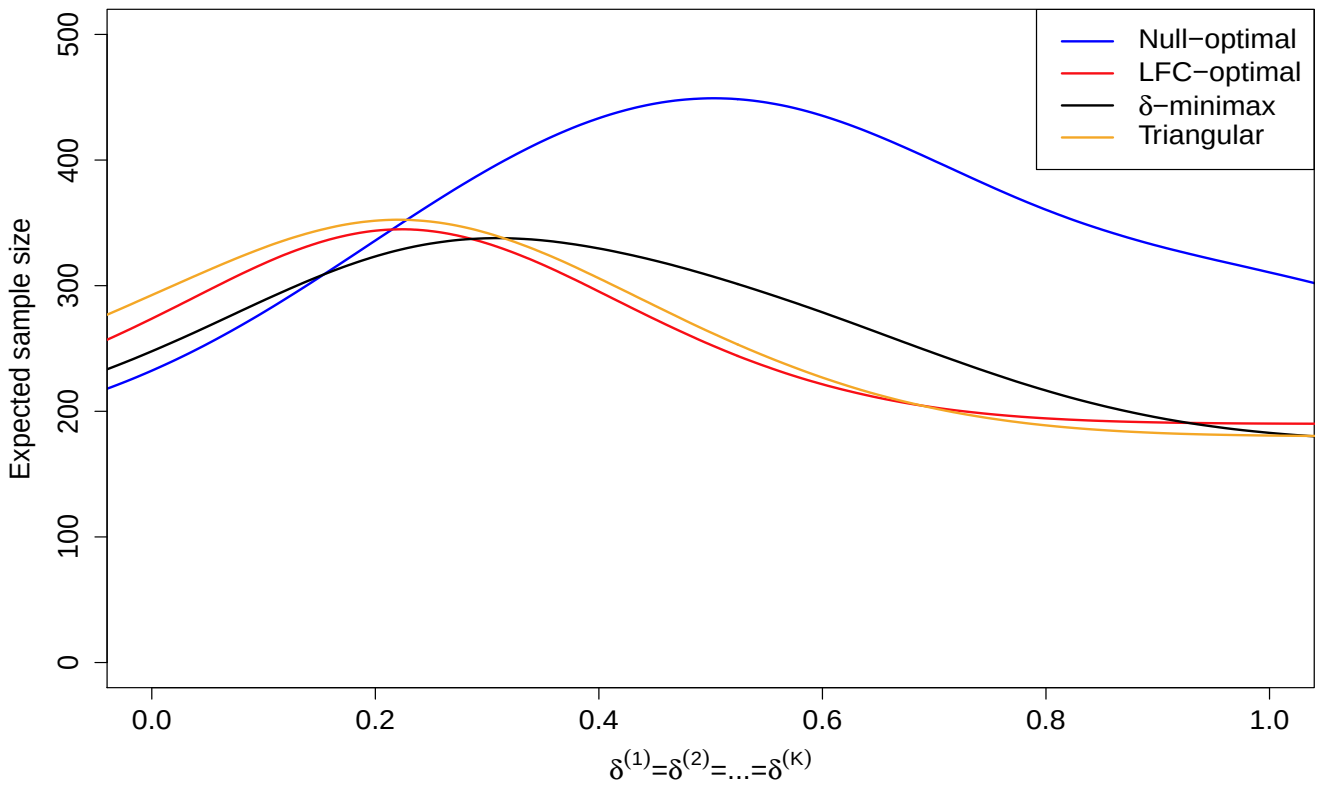
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (2) (number of active treatments = 4, number of stages = 2)



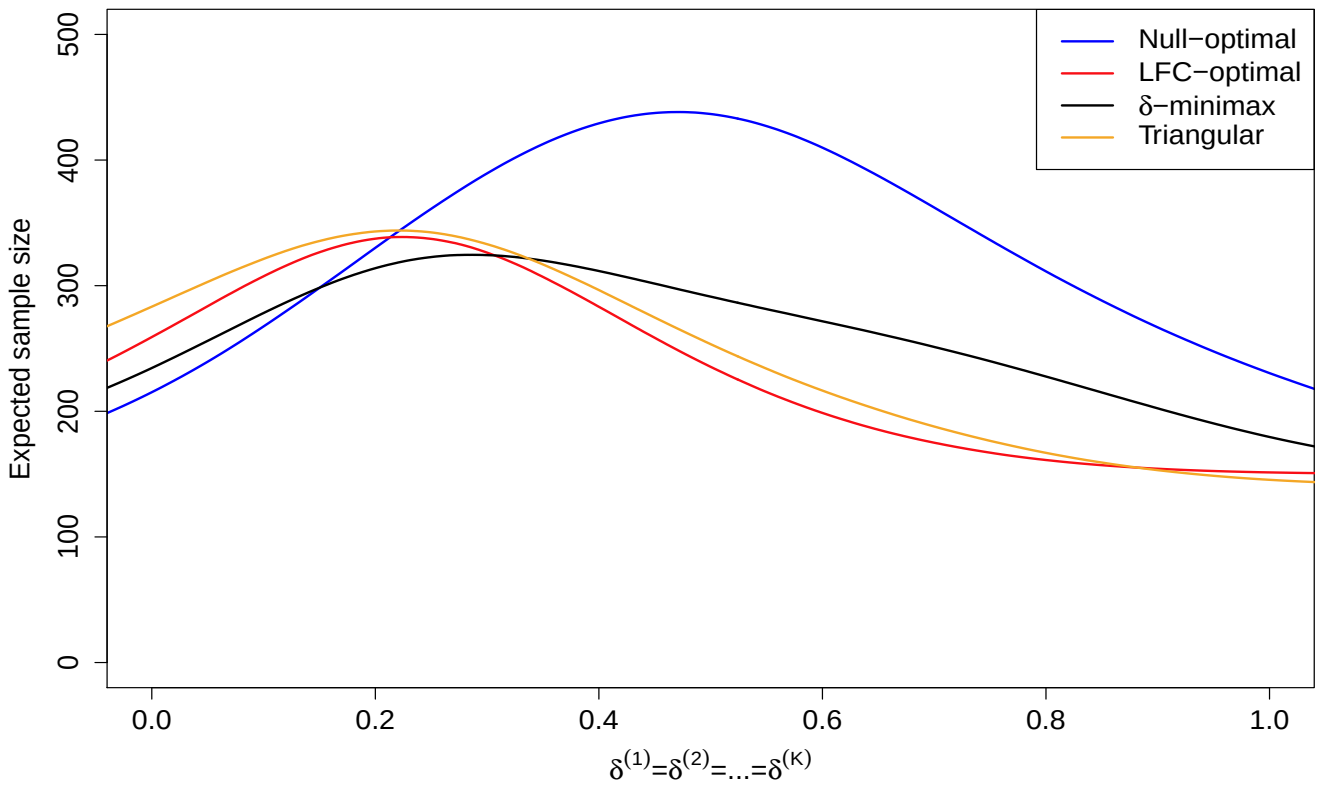
All $\delta^{(k)}$'s varying, and equal.

Expected sample size (2) (number of active treatments = 4, number of stages = 3)



All $\delta^{(k)}$'s varying, and equal.

Expected sample size (2) (number of active treatments = 4, number of stages = 4)

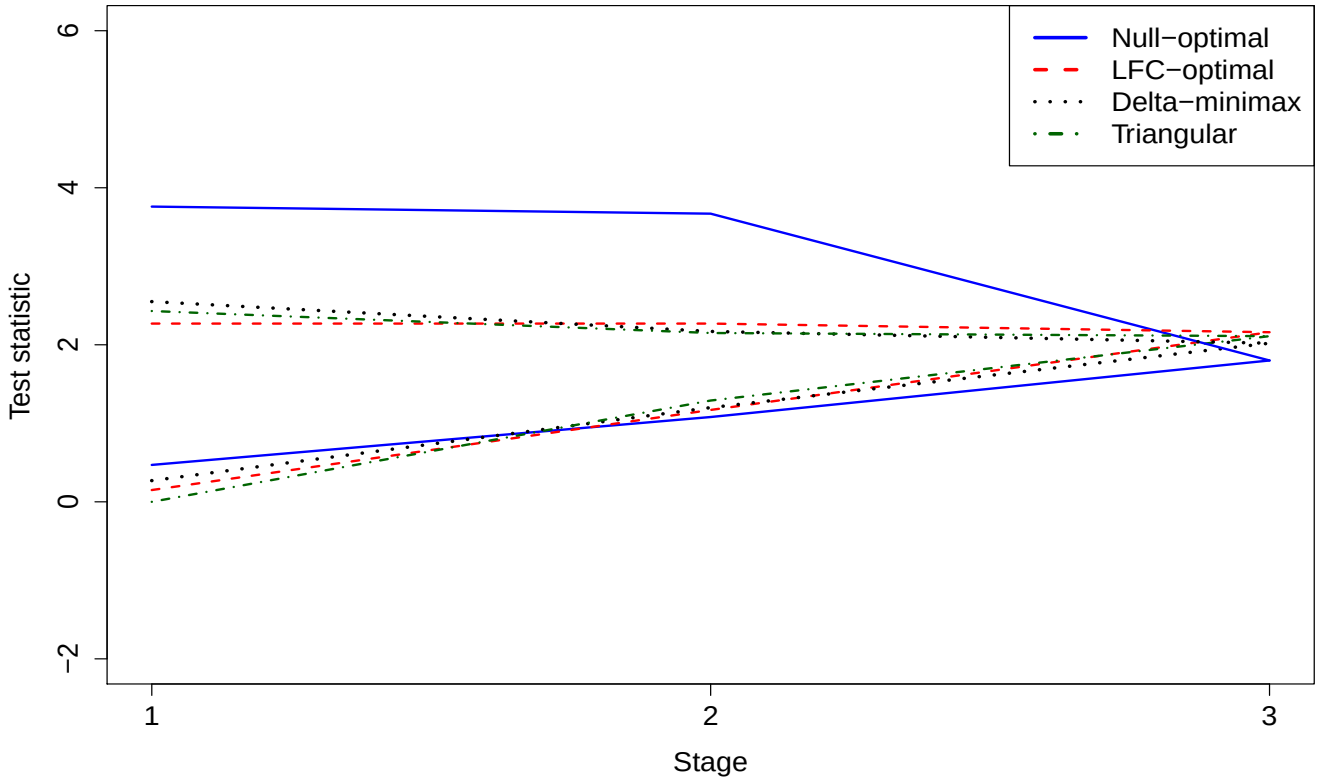


All $\delta^{(k)}$'s varying, and equal.

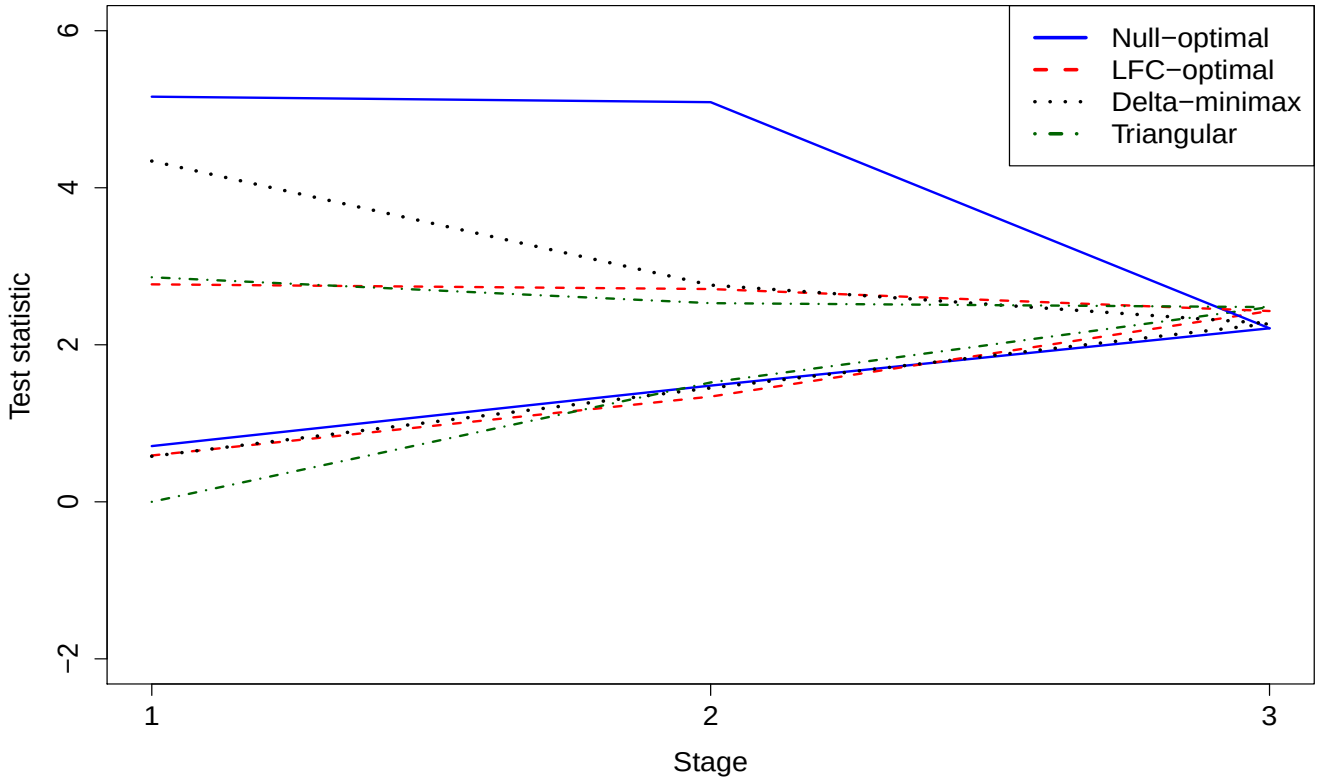
Different numbers of active treatments

- We also looked at $K = 2$, $K = 6$.
- Generally, as number of active arms increases:
  - Differences between the four designs increases.
  - All designs have stricter efficacy boundaries.
  - All designs have higher futility boundaries (i.e. less strict).
  - Performance of group-sequential designs gets worse at criteria considered here. Makes sense as number of ineffective treatments is assumed to increase.

Stopping boundaries of different designs, 2 active arms



Stopping boundaries of different designs, 6 active arms



Optimal allocation ratio

- For multiple active treatments, allocating more patients to the control arm increases the power. For a one-stage trial, it can be shown that the optimal allocation ratio is $\sqrt{K}$.
- No work done for MAMS - STAMPEDE trial ($K = 5$) used allocation ratio of 2 - is this best?
- Simulated annealing process can incorporate allocation ratio as a parameter. Changes correlation between test statistics, so more time consuming.

	Optimal allocation		
Design	K=2	K=4	K=6
H_G-optimal	1.09	1.39	1.72
LFC-optimal	1.12	1.28	1.47
δ-minimax	1.15	1.23	1.45

- However, reduction in expected sample size relatively low (≈ 10 for $K = 4$).
- Could using an adaptive randomisation ratio perform better?

Recommendations

- Generally, a focus on futility boundaries is important - they should be high so treatments that have little chance of being best can be dropped earlier.
- Performance of design depends on true treatment effects. The δ-minimax design performs fairly well in all situations, so is a good design to pick.
- Finding optimal designs is very computationally intensive - if time is limited, the triangular design is much quicker and generally performs well.
- A higher fixed allocation to controls makes little difference to the efficiency, and may in fact put patients off.

Acknowledgements

We are grateful to the MRC Hub for Trials Methodology Research network for funding a research visit. We thank Dominic Magirr for useful discussions and providing code.

Additional slides

Finding optimal designs

- To find optimal design, require:
  1. a method to evaluate type-I error rate, power, and expected sample size for each design.
  2. a method to search for the optimal design with respect to the sample size and stopping boundaries.
- Problems:
  1. Although analytic formulae have been derived to assess operating characteristics, they are time consuming for more than 2 stages.
  2. additionally when there are more than 2 stages, search space is large and consists of many local optima.
- Proposed approach:
  1. Use efficient simulation approach to evaluate operating characteristics. Large number of replicates ($>$250,000). Reduce time by using same simulation replicates to assess different designs.
  2. Use stochastic search technique to search set of possible designs. Simulated annealing used in one arm case also works for MAMS trial.

Analytic method

- Magirr, Jaki and Whitehead derive formulae for the type-I error rate, power, and expected sample size in this case. E.g. type-I error rate:

$$1 - \underbrace{\int_{-\infty}^{\infty} \dots \int_{\infty}^{\infty}}_{\text{J times}} \left[\sum_{j=1}^{J} \Phi_j(L_j(t), U_j(t), \Sigma_j) \right]^K \phi(t_1) \dots \phi(t_J) dt_1 \dots dt_J,$$

J	Time taken to evaluate
2	$<$1 second
3	5.8 minutes
4	8.25 hours

- So if we just need to check one design, then the analytical formulae should be fine for $J \leq 4$.

Finding optimal designs

- To find optimal designs, must search over n, and stopping boundaries. Several methods proposed to find optimal designs.
- For two-stage designs, a grid search technique is feasible, since few parameters.
- For more than two stages, too many parameters. Eales and Jennison (1992) propose using dynamic programming. Does not apply when there is more than one active treatment though.
- Simulated annealing method proposed previously for multi-stage designs with one active treatment. Could be extended, as only requires being able to calculate type-I error rate, power, and expected sample size of designs.
- However, need to evaluate properties for tens of thousands of designs. Using analytic formulae will take too long for $J > 2$.

Efficient simulation method

- For design $(n, f_1, \ldots, f_J, e_1, \ldots, e_J)$, wish to find type-I error rate, power, and expected sample size at some δ.
- Estimate these quantities using simulated data, but as efficiently as possible.
- Want evaluating same design twice to give same answer, so simulate large number of replicates, and re-use.
- Let $X_i = \{x_{ijk},\ j = 1, \ldots, J; k = 1, \ldots, K\}$ be a matrix of random variables with independent rows and such that $(x_{ij1}, \ldots, x_{ijk}) \sim MVN(0, \Sigma)\ \forall\ i, j$; where $\Sigma_{rs} = 1$ if $r = s$ and $\frac{1}{2}$ if $r \neq s$.
- Following relations give $Z = \{Z_{jk}\}$ with same distribution as z-test statistics:

$$Z_1^{(k)} = x_{i1k} + \sqrt{\frac{nr}{2\sigma^2}}\delta^{(k)}$$

$$Z_j^{(k)} = \sqrt{\frac{j-1}{j}}Z_{j-1}^{(k)} + \sqrt{\frac{1}{j}}x_{ijk} + \sqrt{\frac{nr}{2j\sigma^2}}\delta^{(k)}.$$

Possible extension - adaptive allocation ratio

- Previous slide assumed that allocation to controls does not differ as arms are dropped.
- Since optimal allocation ratio depends on number of active arms, ideally it should reduce as arms are dropped.
- Could extend simulation procedure to take this into account. Even more computationally intensive to do however.
- Could there be a way of extending the analytic formulae to take this into account?
- Would increase in efficiency be worth it?