Design of multi-arm trials with and without interim analyses

James Wason
MRC Biostatistics Unit
Cambridge, UK

MRC Hubs for Trials Methodology Research
MRC Biostatistics Unit Hub

Overview

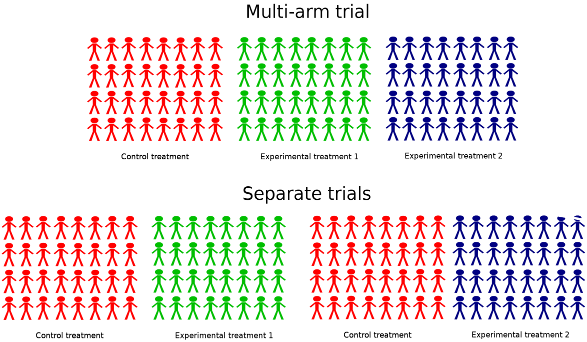
- Multi-arm trials and their advantages.
- Multi-arm multi-stage trials (MAMS):
    1. Group-sequential MAMS;
    2. Drop-the-loser designs;
    3. Adaptive randomization.

Randomised controlled trials

- The traditional way to assess the effectiveness of a new experimental treatment is to use a randomised controlled trial.
- Patients are randomised between the experimental treatment and a control treatment.
- A statistical test is done after all patients are assessed for response to treatment to test the null hypothesis that the experimental treatment is equal to or worse than the control treatment.
- This has been considered the gold-standard approach for many years, and has worked well.

Multi-arm trials

- Multi-arm trials are of great current interest as they considerably reduce sample size requirements when testing multiple new treatments.



Issues in multi-arm trials

- Although more efficient than separate trials, there are some additional issues in multi-arm trials.

Issues in multi-arm trials: 1) which arms to compare?

- It is not obvious which comparisons should be carried out following a multi-arm trial. Possibilities:
  1. Compare each experimental arm vs control separately.
  2. Compare all possible pairs of arms.
  3. Omnibus test - compare whether all treatments have the same mean.
- 2) has been suggested previously, but it is unlikely that there will be sufficient power to detect differences between experimental treatments.
- 3) may have higher power than 1) when there are several experimental treatments better than the control; however, may miss an effective treatment if 1-2 experimental treatments are better than control.
- Since most new treatments are not better than control, 1) tends to be suggested most.

Issues in multi-arm trials: 2) control of type-I error rate

- In a typical multi-arm trial, there are multiple null hypotheses being tested.
- Take the example of comparing several experimental treatments against a common control. There are two schools of thought on whether this should be accounted for.
  - A multi-arm trial can be thought of as separate trials within the same protocol - no correction for multiple testing made for separate trials, so no need for it here.
  - The maximum probability of making a type-I error (the family-wise error rate, FWER) needs to be controlled at a specified level.

Issues in multi-arm trials: 2) control of type-I error rate

- Depending on the context of the trial I tend to think that a multiple testing correction should be used. Otherwise I think the FWER should at least be stated.
- For a multi-arm trial, the maximum family-wise error rate occurs when all experimental treatments have the same effect as the control treatment.
- Thus, it is straightforward to specify the maximum FWER.

Issues in multi-arm trials: 2) control of type-I error rate

- The effect of multiple testing correction on the power is fairly large, but still multi-arm trials provide increased efficiency.
- For example, for a controlled phase II trial with a total of 50 patients per arm:

Number experimental treatments	Sample size (separate)	Sample size (multi-arm, no adjustment)	Sample size (multi-arm, full adjustment)
2	200	150	180
3	300	200	256
4	400	250	335
5	500	300	420

Issues in multi-arm trials: 3) powering the trial

- In an RCT, the power is done by specifying a clinically relevant difference (CRD) such that the trial has a high chance to reject the null hypothesis when the true difference is equal to the CRD.
- Less obvious how to power a multi-arm trial.
- It could be powered to detect the best treatment, or just to detect any that have a certain treatment effect.
- One possibility is to use the ‘least favourable configuration (LFC)’. This assumes that one treatment is effective, with CRD treatment effect; the other treatments are assumed to have a treatment effect below a pre-specified ‘uninteresting treatment effect’. The power is then the probability of recommending the effective treatment.

Multi-arm multi-stage trials

- Can also add interim analyses to a multi-arm trial (=multi-arm multi-stage).
- Allows information gathered on treatment effectiveness during the trial to change the design.
- For example, if a certain experimental treatment is performing poorly, could drop it from the trial, or reduce future allocation of patients to that treatment.
- Could also can allow early stopping for efficacy if an effective treatment is found.
- Interim analyses add additional efficiency and also make the trial more ethical.

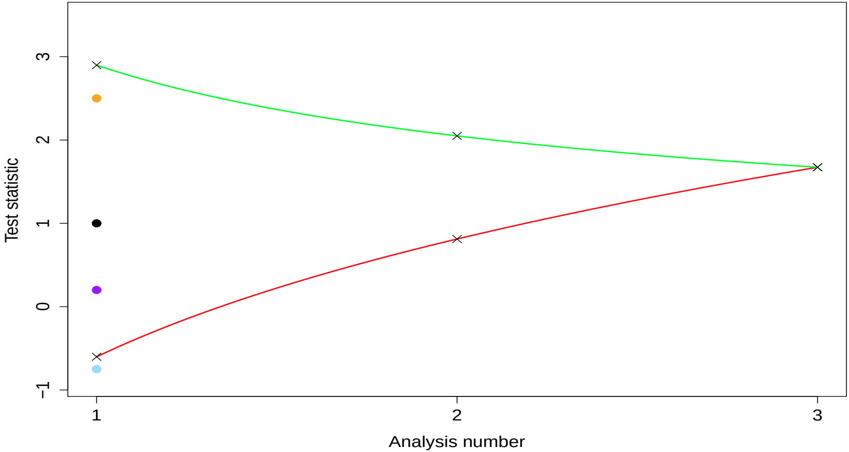
Multi-arm multi-stage trials

- There are several methods for using interim analyses to improve the efficiency of a multi-arm trial.
- I will cover three distinct methods:
  1. Group-sequential designs;
  2. Drop-the-loser designs;
  3. Adaptive randomisation.
- Firstly, group-sequential multi-arm trials.

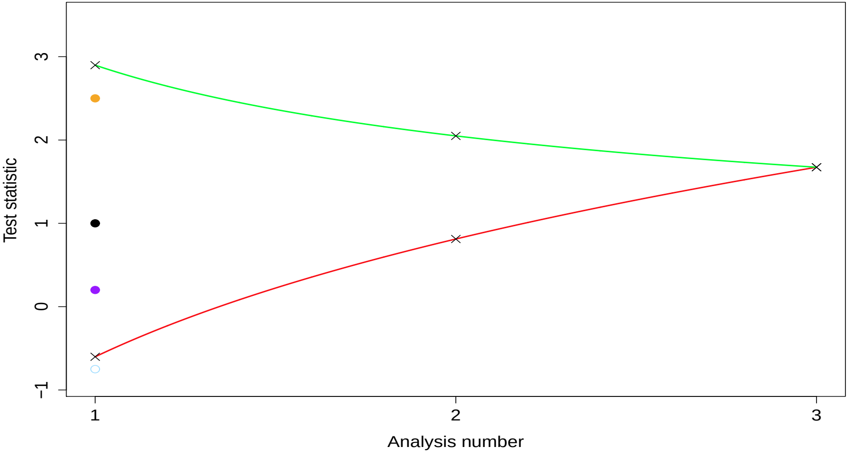
Group-sequential MAMS trials

- Dunnett (1955) developed method for testing multiple experimental treatments against a control with exact correction for multiple testing.
- Methodology extended to multi-arm trials with interim analyses (Magirr, Jaki and Whitehead, 2012).
- Binding futility and efficacy stopping boundaries are specified.
- At each interim analysis, test statistics testing each experimental treatment against control are calculated.
    - If below the futility boundary, that treatment is dropped for futility.
    - If above the efficacy boundary, the trial stops with the conclusion that the treatment is effective.

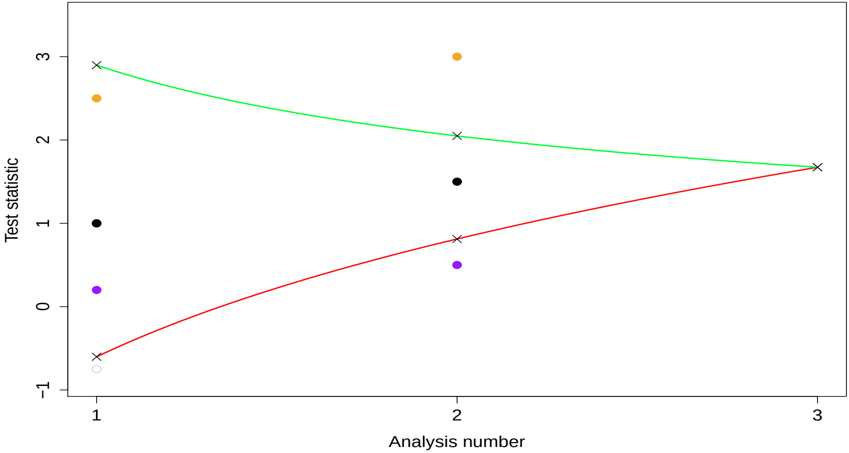
Group-sequential MAMS trials



Group-sequential MAMS trials



Group-sequential MAMS trials



Choosing stopping boundaries

- The choice of boundaries is an important problem. Together with the sample size recruited per arm, per stage, they affect the operating characteristics of the design.
- Given a target FWER and power, there are infinitely many boundary shapes.
- Need a further way of distinguishing designs.
- Often in group-sequential trials, the expected sample size (ESS) is of interest.

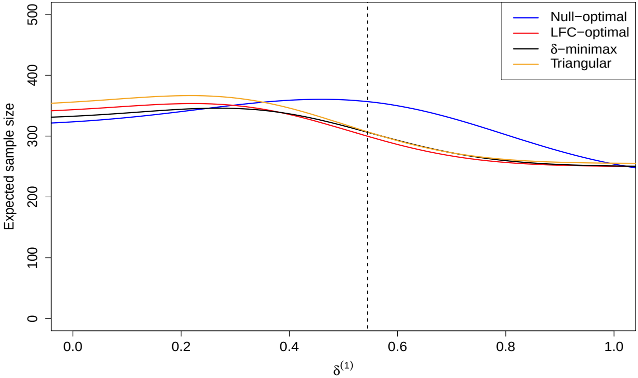
Expected sample size

- The ESS is the average sample size used if the trial were repeated many times.
- Denote $\delta^{(k)}$ as the difference in treatment effect between arm k and the control arm, $\delta = (\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(K)}$.
- Then ESS depends on δ.
  1. For very low values of δ, the trial is very likely to stop early for futility;
  2. For high values of δ, the trial is very likely to stop early for efficacy.

Optimal designs

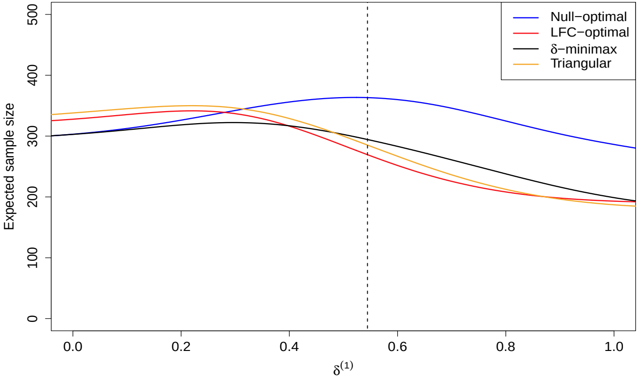
- This leads to the idea of an *optimal* design.
- An optimal design meets the FWER and power constraints, and minimises the ESS at some value of δ.
- Infinite number of optimality criteria, so we restrict ourselves to three:
  - Global-null-optimal design: optimal for $\delta^{(1)} = \delta^{(2)} = \ldots = \delta^{(k)} = 0$.
  - LFC-optimal design: optimal under the least favourable configuration.
  - δ-minimax design - aims to minimise the maximum expected sample size.
- Much more detail given in Wason and Jaki (2012, Statistics in Medicine).

Expected sample size (1)



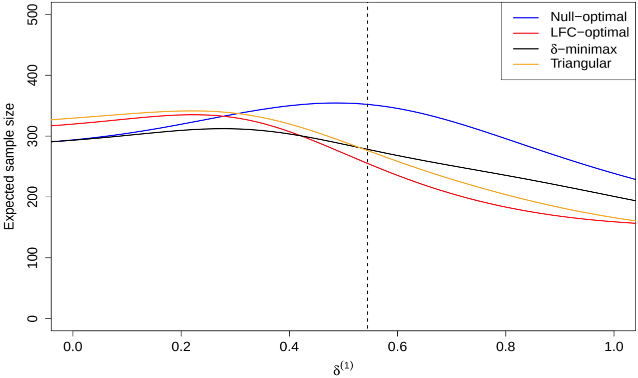
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (1)



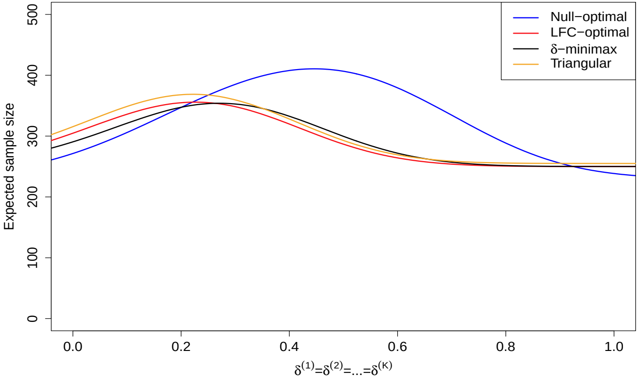
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (1)



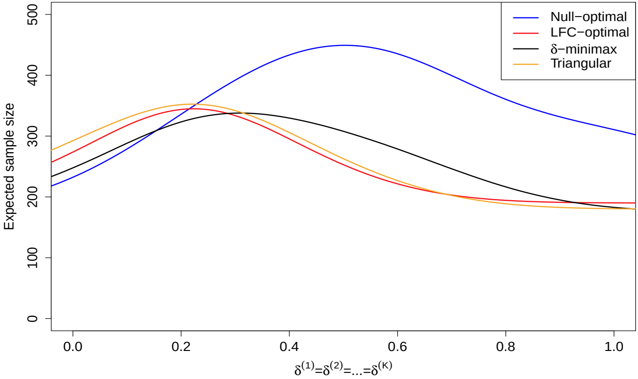
Only $\delta^{(1)}$ varying, all others $= \delta_0$.

Expected sample size (2)



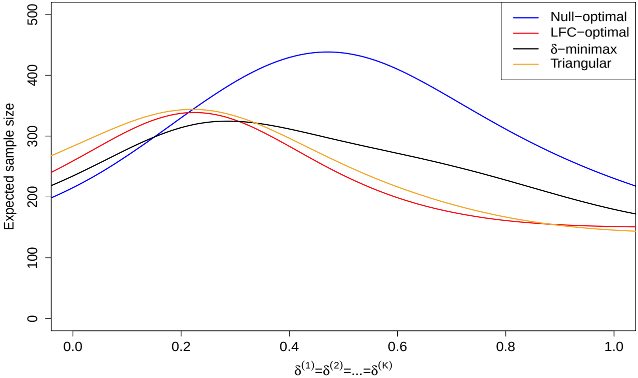
All $\delta^{(k)}$'s varying, and equal.

Expected sample size (2)



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

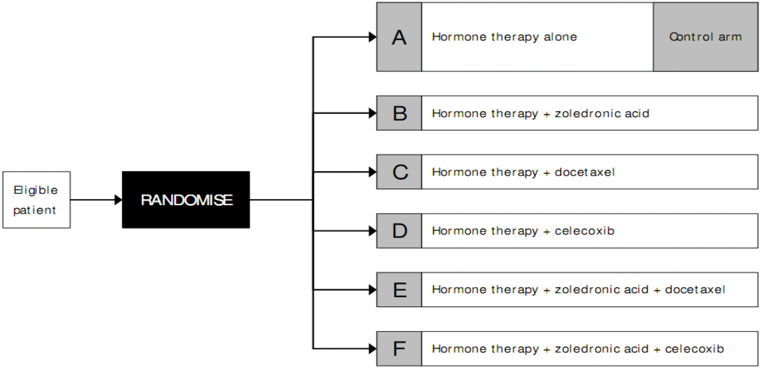
Expected sample size (2)



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

Example: MRC STAMPEDE trial

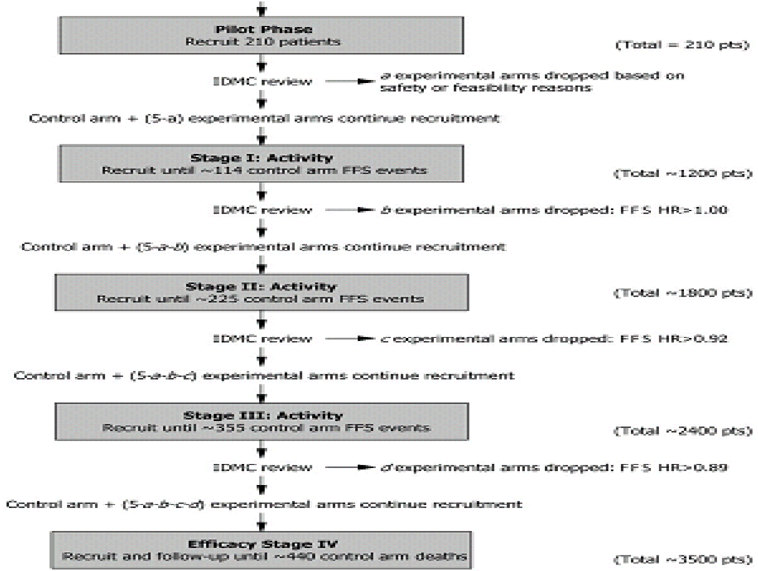
- The MRC STAMPEDE trial is a MAMS trial currently running in men with hormone-sensitive advanced prostate cancer.
- Five stages and six arms.
- Following figures taken from Sydes et al.



Eligible patient
RANDOMISE
A
Hormone therapy alone
Control arm
B
Hormone therapy + zoledronic acid
C
Hormone therapy + docetaxel
D
Hormone therapy + celecoxib
E
Hormone therapy + zoledronic acid + docetaxel
F
Hormone therapy + zoledronic acid + celecoxib

Figure 1
STAMPEDE trial arms. The randomisation ratio is 2A : 1B : 1C : 1D : 1E : 1F. HT = hormone therapy, *bid* = twice daily.

Example: MRC STAMPEDE trial



Pilot Phase
Recruit 210 patients

(Total = 210 pts)

IDMC review → *a* experimental arms dropped based on safety or feasibility reasons

Control arm + (5-a) experimental arms continue recruitment

Stage I: Activity
Recruit until ~114 control arm FFS events

(Total ~1200 pts)

IDMC review → *b* experimental arms dropped: FFS HR>1.00

Control arm + (5-*a*-*b*) experimental arms continue recruitment

Stage II: Activity
Recruit until ~225 control arm FFS events

(Total ~1800 pts)

IDMC review → *c* experimental arms dropped: FFS HR>0.92

Control arm + (5-*a*-*b*-*c*) experimental arms continue recruitment

Stage III: Activity
Recruit until ~355 control arm FFS events

(Total ~2400 pts)

IDMC review → *d* experimental arms dropped: FFS HR>0.89

Control arm + (5-*a*-*b*-*c*-*d*) experimental arms continue recruitment

Efficacy Stage IV
Recruit and follow-up until ~440 control arm deaths

(Total ~3500 pts)

Optimal allocation ratio

- For multiple experimental 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 (five stages, five experimental arms) used allocation ratio of 2 - is this best?

	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 ($\approx$ 10 for four experimental arms).

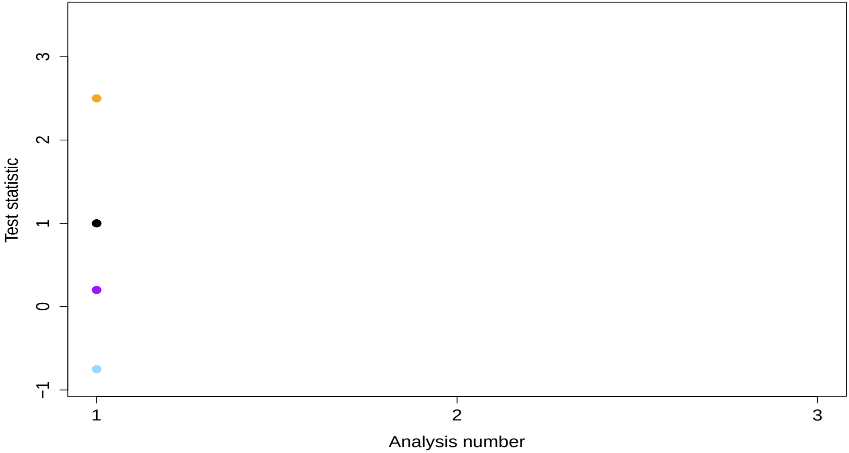
Drop-the-loser designs

- Expected sample size $= 292.5$, but 95% quantile is 432.
- This creates problems - how much funding should such a trial apply for? It cannot be the expected sample size as often that won't be enough.
- Applying for the maximum will mean the funder is committing a lot of money (although may get some back).
- Other issues with logistics, e.g. length of trial staff contracts.
- Although one may lose some efficiency on average, a MAMS design that has a fixed sample size might well be of interest.

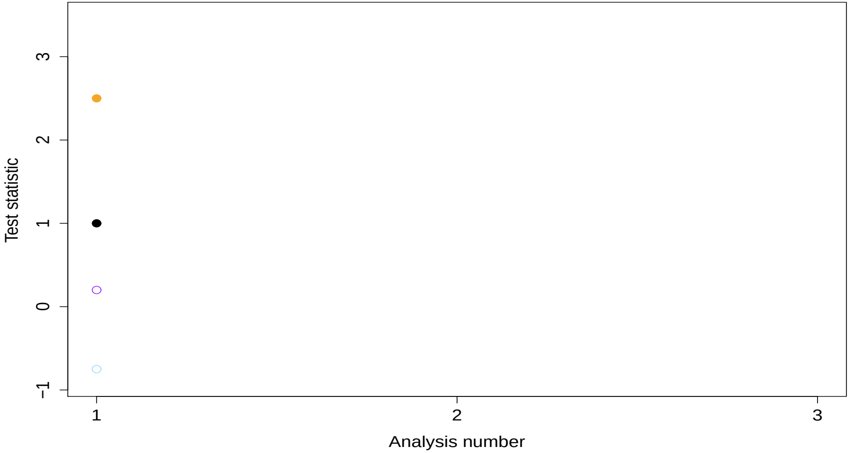
Drop-the-loser designs

- Instead of setting futility and efficacy thresholds on test statistics, one can set the number of treatments to progress at each interim analysis.
- The treatments that progress are the ones with the highest test statistics at the interim.
- The main advantage of this approach is that the sample size is fully known in advance (no variability).
- For example, a 4:2:1 design:

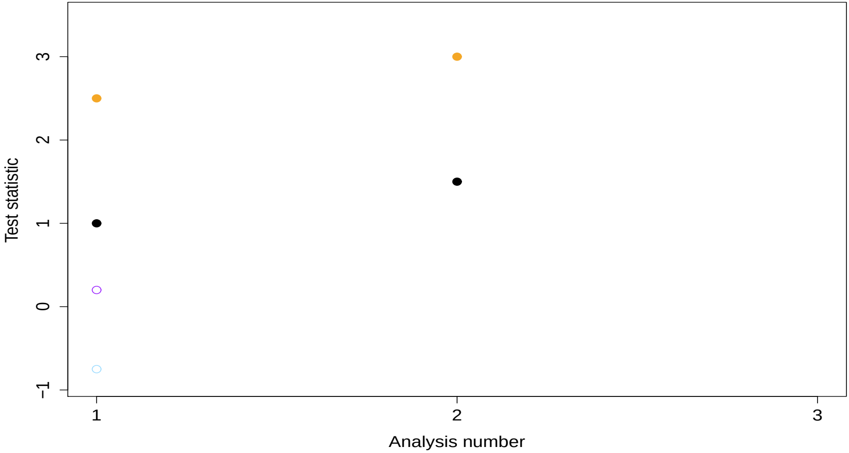
Multi-stage drop-the-loser designs



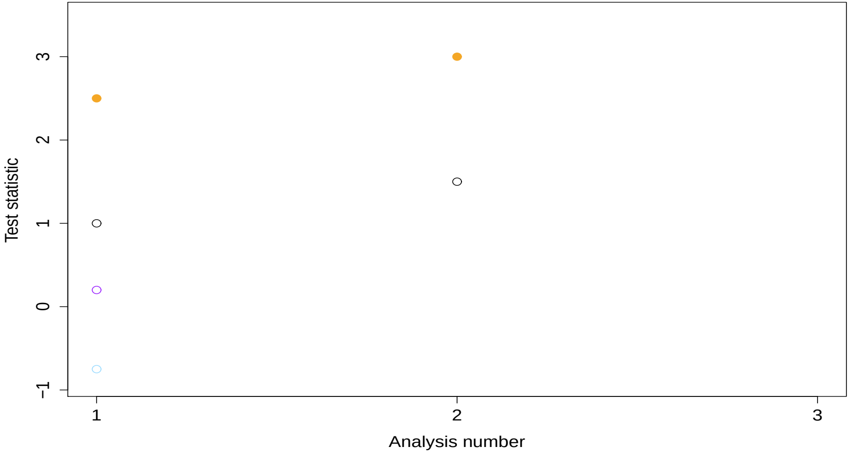
Multi-stage drop-the-loser designs



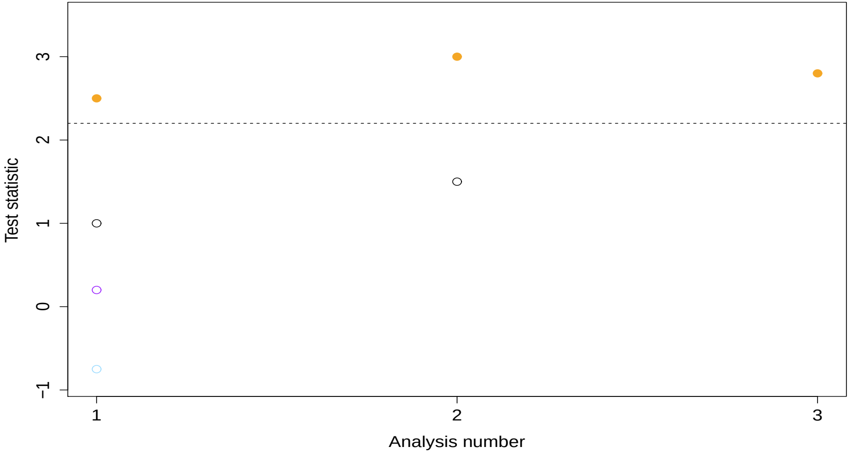
Multi-stage drop-the-loser designs



Multi-stage drop-the-loser designs

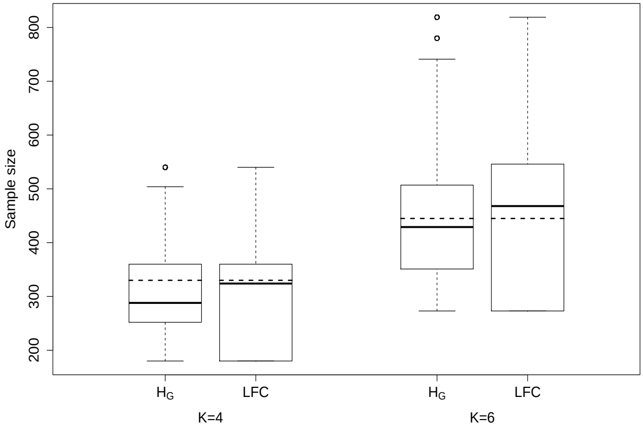


Multi-stage drop-the-loser designs



Comparison of sample size requirements

- From some submitted work, a comparison of the sample size requirements of a drop-the-loser design to that of a group-sequential MAMS design:



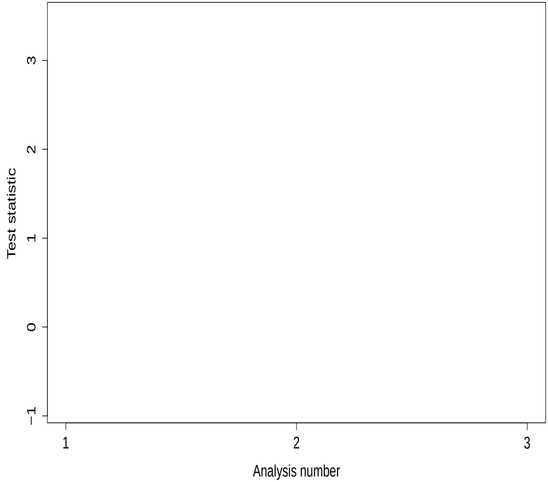
Advantages and disadvantages

- The main advantage of a drop-the-loser design is that it has a fixed sample size which is generally close to the median of the group-sequential design.
- This fixed sample size should make it easier to plan studies.
- Disadvantages include the inflexibility of the decision rules. For example, if one treatment is very close in effect to another, then it might be that only one can go through. Likewise, if all experimental treatments are performing poorly, this design does not have scope for stopping early.
- Deviating from the design may cause incorrect type-I error and power.
- A hybrid design may be more suitable (Stallard and Friede).

Adaptive randomisation

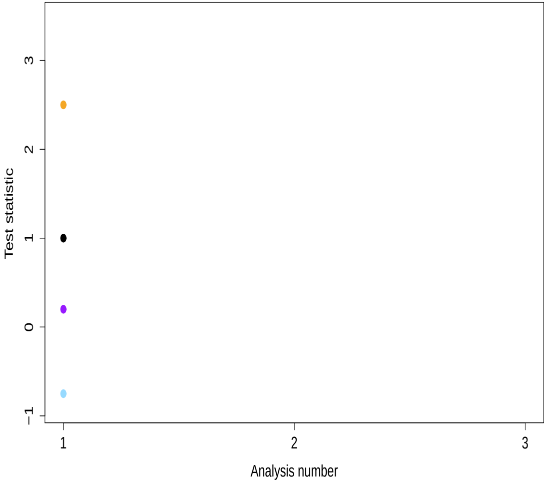
- Outcome-adaptive randomisation in essence is the idea that the allocation of patients should be increased to successful treatments and decreased to unsuccessful treatments.
- There is a lot of literature on its use in two-arm trials.
- Consensus is that it can lead to better average patient outcomes, but lower statistical power.
- When there are multiple experimental arms, the allocation to the control group set separately, and AR applied just to the experimental arms.
- This maintains the statistical power.

Adaptive randomisation



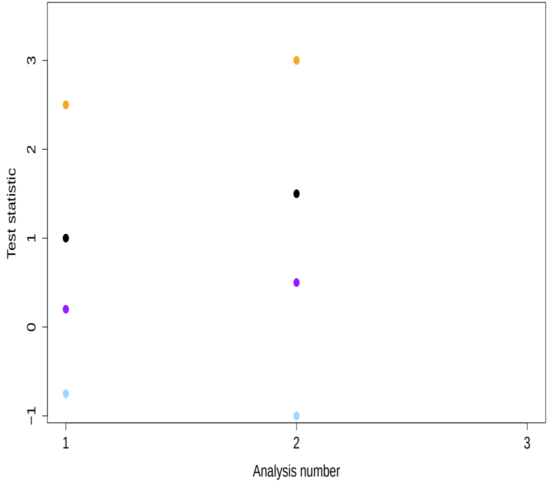
Arm	Allocation
Control	0.2
Orange	0.2
Black	0.2
Purple	0.2
Blue	0.2

Adaptive randomisation



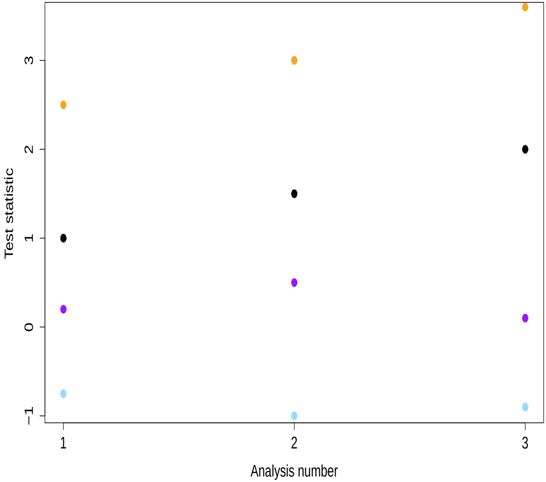
Arm	Allocation
Control	0.3
Orange	0.3
Black	0.2
Purple	0.15
Blue	0.05

Adaptive randomisation



Arm	Allocation
Control	0.35
Orange	0.35
Black	0.175
Purple	0.1
Blue	0.025

Adaptive randomisation

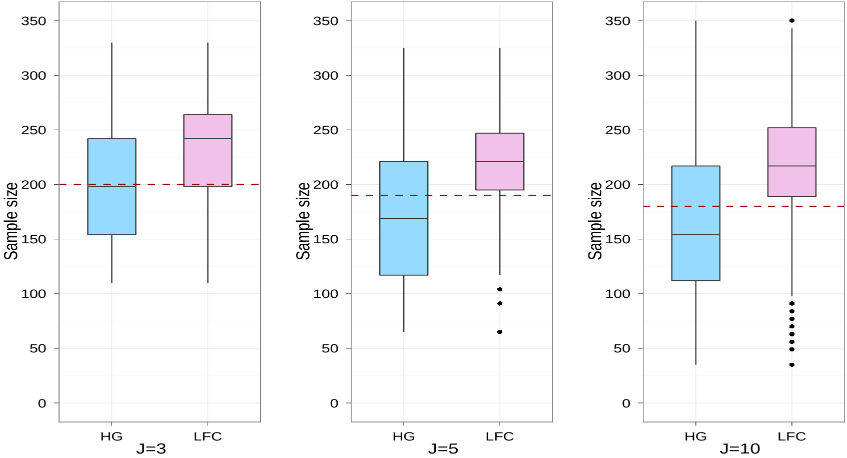


Arm	Allocation
Control	0.35
Orange	0.35
Black	0.175
Purple	0.1
Blue	0.025

Comparison of AR and MAMS

- A well calibrated AR procedure results in fairly similar statistical properties to a group-sequential MAMS procedure.
- Generally AR is more efficient when there are effective experimental treatments, and less efficient when none are effective.
- AR will probably have more problems when patient characteristics or treatment effects change as the trial progresses.
- However, AR is much more flexible when it comes to dealing with delay, missing data, intermediate outcomes, or patient subgroups.

Comparison of designs



Blue boxplots - sample size used when all experimental treatments are ineffective.
Pink boxplots - sample size used when one experimental treatment is effective.
Dashed line - sample size required by AR.

Conclusion

- Multi-arm trials provide a big gain in efficiency over separate trials
- Several designs with interim analyses to choose from (group-sequential, drop-the-loser, and adaptive randomisation). Generally statistical performance is fairly similar, but each has its advantages and disadvantages.
- All designs become less efficient when there is a long delay between recruiting patients and assessing their treatment response. An intermediate endpoint may be useful for aiding early decision making.

References

C.W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50:1096-1121, 1955.

D. Magirr, T. Jaki, and J. Whitehead. A generalized Dunnett test for multiarm multistage clinical studies with treatment selection. Biometrika 99:494-501, 2012.

N. Stallard and T. Friede. A group-sequential design for clinical trials with treatment

J. Wason, D. Magirr, M. Law, T. Jaki. Some recommendations for multi-arm multi-stage trials. Statistical Methods in Medical Research Epub.

J. Wason and T. Jaki. Optimal design of multi-arm multi-stage trials. Statistics in Medicine 31:4269-4279

Formal description of Adaptive randomisation

- More formally, specify prior distributions for effectiveness of each treatment.
- Let there be J stages, with a total of N patients recruited, equally distributed between the stages.
- After each stage, calculate posterior probabilities $\mathbb{P}(\delta_k > 0|data)$ for each arm.
- Allocation to arm k, π_k, in the next stage is set proportional to these probabilities.

$$\pi_k \propto \begin{cases} \dfrac{\mathbb{P}\Big(\delta_k > 0|\, Y_{jk}, Y_{j0}, n_{jk}, n_{j0}\Big)^{\gamma(j/J)}}{\displaystyle\sum_{i=1}^{K} \mathbb{P}\Big(\delta_i > 0|\, Y_{ji}, Y_{j0}, n_{ji}, n_{j0}\Big)^{\gamma(j/J)}} & \text{if k=1, } \ldots \text{, K;} \\ \frac{1}{K} \exp\Big(\max(n'_1, n'_2, \ldots, n'_K) - n'_0\Big)^{\eta(n'/N)} & \text{if k=0,} \end{cases} \tag{1}$$

where γ and η are increasing functions.

Formal description of Adaptive randomisation

- At the end of the trial, test statistics for the effect of each arm against control, $(Z_1, \ldots, Z_K)$ are calculated.
- If Z_k is above a critical value c, then reject $H_0^{(k)}$.
- N and c are chosen to control the FWER/type-I error rate and power of the design.
- The choice of $\gamma()$ affects the efficiency of the design substantially.
- We have explored a large number of possibilities and found the most efficient choice (results in submitted paper, available upon request).