Flexible Sequential Designs for Multi-Arm Clinical Trials

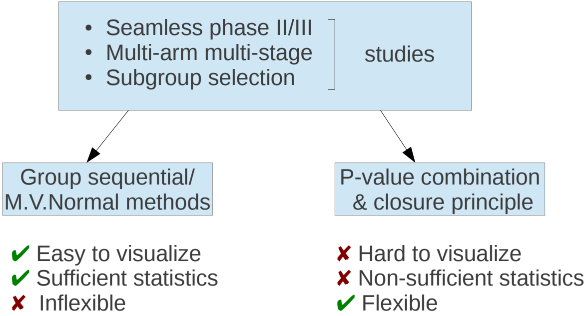
Dominic Magirr*, Nigel Stallard**, Thomas Jaki*

*Department of Mathematics and Statistics, Lancaster University
**Warwick Medical School, University of Warwick

June 2013

Outline

Purpose of talk



A multi-arm phase II trial

Wilkinson & Murray, 2001; Stallard & Todd, 2003.

Objective: "To investigate whether **Galantamine** significantly improves the core symptoms of **Alzheimer's disease**".

Treatment: Galantamine 18 mg/day, 24 mg/day, 36 mg/day vs. Placebo

Endpoint: Change in Alzheimer's Disease Assessment Scale (assumed to be normally distributed) after 3 months of treatment.

Hypothesis testing

In this case there are 3 null hypotheses of interest:

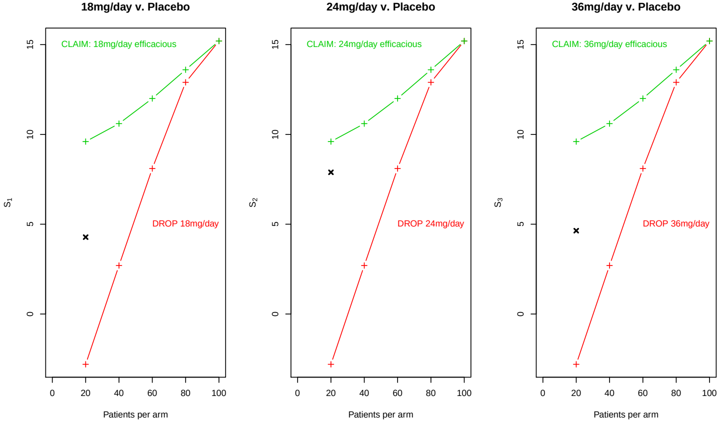
$$H_1 : \theta_1 \leq 0,$$
$$H_2 : \theta_2 \leq 0,$$
$$H_3 : \theta_3 \leq 0,$$

where θ_k is the treatment effect of dose $k = 1, 2, 3$.

H_k will be rejected if $S_k = (\bar{X}_k - \bar{X}_0)n/2\sigma^2$ is “large enough”.

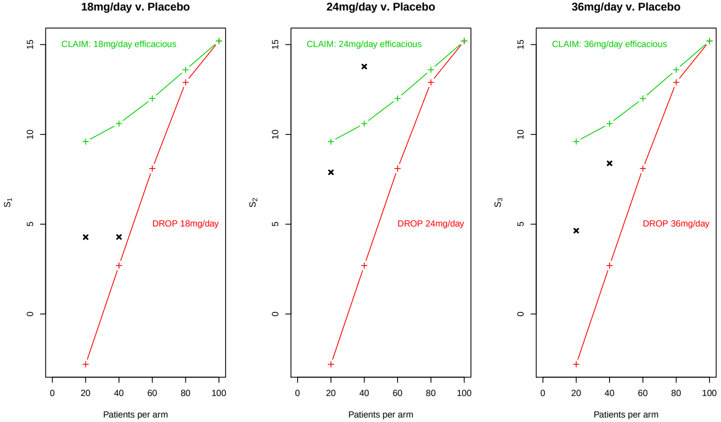
Statistical monitoring

First interim analysis



Statistical monitoring

Second interim analysis

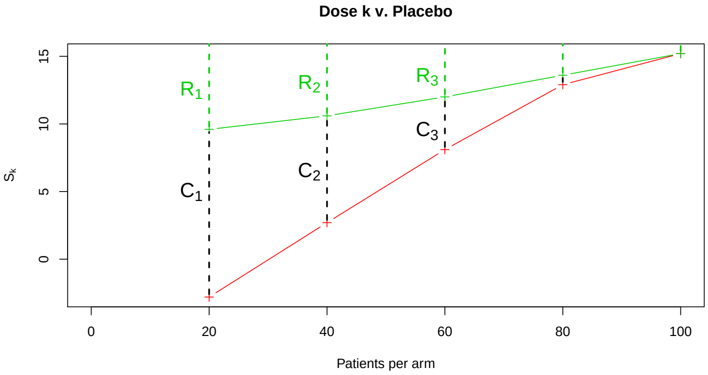


Statistical monitoring

Second interim analysis

- 36mg/day arm was dropped due to safety concerns.
- Recruitment to 18mg/day arm was continued (the lower boundary was later crossed at the 4th interim analysis).
- It was concluded that 24mg/day was **safe** and **effective**.
- Is this conclusion justified?

Type I error probabilities



1. What is the probability that S_k crosses the upper boundary?
2. What is the probability that $\max_{k=1,2,3} S_k$ crosses the upper boundary?

Per-comparison error rate

$$P_0\left\{S_k^{(1)} \in R_1\right\} = 0.001$$

$$P_0\left\{S_k^{(1)} \in C_1\right\} \cap \left\{S_k^{(2)} \in R_2\right\} = 0.008$$

$$\vdots$$

$$P_0\left\{S_k^{(1)}, \ldots, S_k^{(4)} \in C_1 \times \cdots \times C_4\right\} \cap \left\{S_k^{(5)} \in R_5\right\} = 0.0005$$

$$P_0 \bigcup_j \left\{S_k^{(1)}, \ldots, S_k^{(j-1)} \in C_1 \times \cdots \times C_{j-1}\right\} \cap \left\{S_k^{(j)} \in R_j\right\} = 0.025$$

Familywise error rate

$$P_0\left\{S_k^{(1)} \in R_1\right\} = 0.001$$

$$P_0\left\{S_k^{(1)} \in C_1\right\} \cap \left\{S_k^{(2)} \in R_2\right\} = 0.008$$

$$\vdots$$

$$P_0\left\{S_k^{(1)}, \dots, S_k^{(4)} \in C_1 \times \cdots \times C_4\right\} \cap \left\{S_k^{(5)} \in R_5\right\} = 0.0005$$

$$P_0 \bigcup_j \left\{S_k^{(1)}, \dots, S_k^{(j-1)} \in C_1 \times \cdots \times C_{j-1}\right\} \cap \left\{S_k^{(j)} \in R_j\right\} = 0.025$$

$$P_0 \bigcup_{k=1}^{3} \bigcup_j \left\{S_k^{(1)}, \dots, S_k^{(j-1)} \in C_1 \times \cdots \times C_{j-1}\right\} \cap \left\{S_k^{(j)} \in R_j\right\} = 0.063$$

Control of familywise error rate

To control the FWER at level α, one could simply increase the upper stopping boundary, i.e., solve the following equations for $u_1, \ldots, u_5$.

$$P_0 \bigcup_k \left\{ S_k^{(1)} \in R_1 \right\} = \alpha_1^*$$

$$P_0 \bigcup_k \bigcup_{j=1}^{2} \left\{ S_k^{(1)}, \ldots, S_k^{(j-1)} \in C_1 \times \cdots \times C_{j-1} \right\} \cap \left\{ S_k^{(j)} \in R_j \right\} = \alpha_2^*$$

$$\vdots$$

$$P_0 \bigcup_k \bigcup_{j=1}^{5} \left\{ S_k^{(1)}, \ldots, S_k^{(j-1)} \in C_1 \times \cdots \times C_{j-1} \right\} \cap \left\{ S_k^{(j)} \in R_j \right\} = \alpha_5^*$$

where $\alpha_1^* \leq \cdots \leq \alpha_5^* = \alpha$.

See *Follmann et al.,1994; Chen et al., 2010; Magirr, Jaki & Whitehead (MJW), 2012.*

Alternative group-sequential/treatment selection design

Stallard & Todd (ST), 2003

- Let $M = \left\{k : S_k^{(1)} = \max_{k'} S_{k'}^{(1)}\right\}$, i.e., "the best treatment at first interim analysis".
- Solve the following equations for $u_1, \ldots, u_5$.

$$P_0\left\{S_M^{(1)} \in R_1\right\} = \alpha_1^*$$

$$P_0\left\{S_M^{(1)} \in C_1\right\} \cap \left\{S_M^{(2)} \in R_2\right\} = \alpha_2^* - \alpha_1^*$$

$$\vdots$$

$$P_0\left\{S_M^{(1)}, \ldots, S_M^{(4)} \in C_1 \times \cdots \times C_4\right\} \cap \left\{S_M^{(5)} \in R_5\right\} = \alpha_5^* - \alpha_4^*$$

Power and sample size

- The Galantamine trial was powered such that

$$P_{\theta_k=0.5\sigma}(\text{“}S_k \text{ crosses upper boundary”}) = 0.9.$$

- This takes no account of multiple comparisons.
- However, there is no obviously better definition of power in a multi-arm trial.
- One possibility (Dunnett, 1984) is to consider a **least favourable configuration** of treatment effects.

Least favourable configuration

Dunnett, 1984

Need to consider two effect sizes:

1. δ, the smallest clinically relevant improvement (standard design question).
2. $0 \leq \delta_0 < \delta$, the largest marginal improvement such that if $\theta_k = \delta_0$ we would prefer not to proceed further in investigation of treatment k.

- $(\delta_0,\ \delta)$ is 'zone of indifference'.

- 'Least favourable configuration' (LFC): $\theta_1 = \delta$, $\theta_k = \delta_0$ for $k = 2, 3, \ldots$.

Sample size based on LFC

Choose n such that

$$P(\text{“select treatment 1”} \mid \text{LFC}) = 1 - \beta,$$

where

$$\text{“select treatment 1”} \equiv \text{“}S_1 \text{ crosses upper stopping boundary”} \\ (\text{before } S_2, S_3, \ldots)$$

Summary of group-sequential multi-arm trials

1. Require relatively simple (if somewhat tedious) calculations to set up.
2. Once stopping boundaries and sample size are found $\rightarrow$ monitoring the trial is straightforward.
3. All decisions are based on the sufficient statistics $S_k = (\bar{X}_k - \bar{X}_0)n/2\sigma^2$.
4. Familywise error rate is controlled "in the strong sense",

$$\sup_{\theta} P_{\theta}\left\{\text{reject one (or more) true null hypothesis}\right\} \leq \alpha$$

5. Major disadvantage: Lack of flexibility $\rightarrow$ think of what happened in Galantamine trial.

Flexible design methodology

- P-value combination functions. (Bauer & Köhne, 1994)
- Conditional error rate. (Proschan & Hunsburger, 1995)
- Closure principle. (Marcus et al., 1976)

Combining these techniques produces very flexible treatment (or subgroup) selection phase II/III designs. E.g.

- Posch et al., 2005.
- König et al., 2008.

Flexible design methodology

MAIN IDEA: the second-stage design is not pre-specified.

ARCHETYPE: reject $H_0 : \theta = 0$ in favour of $H_a : \theta > 0$ if

$$Z = \sqrt{\frac{n_1}{n_1 + n_2}} Z_1 + \sqrt{\frac{n_2}{n_1 + n_2}} \Phi^{-1}(1 - p_2) \geq 1.96,$$

where, under H_0,

- $Z_1 \sim N(0, 1)$.
- $p_2 \sim U(0, 1)$, irrespective of first-stage data and choice of second stage test.
- n_1 and n_2 are planned first- and second-stage sample sizes, respectively.

Danger of using non-sufficient statistic

Burman & Sonesson, 2006

- Suppose $n_1 + n_2 = 1000$ experimental units are to be recruited.
- This gives power 0.8 if $\theta = 0.08$ and $\sigma = 1$.
- After $n_1 = 100$ observations, it is decided to take an interim look at the data.
- Disappointingly, the observed average effect is slightly negative, $\hat{\theta} = -0.03$.

Danger of using non-sufficient statistic

Burman & Sonesson, 2006

- The experimenter doesn't consider it worthwhile to continue to collect 900 more observations, as planned.
- Instead, the experimenter collects **one** additional observation.
- If X_{101} happens to be, say, 2.5,

$$Z = \sqrt{0.1}(-0.3) + \sqrt{0.9}(2.5) \approx 2.28.$$

  - It is concluded that $\theta > 0$.
  - However, $\hat{\theta} \approx -0.005$.

Danger of using non-sufficient statistic

- This is an extreme (ridiculous) example.
- Nevertheless, it captures the essence of more subtle investigations into the use of non-sufficient statistics in adaptive designs.
- See *Tsiatis & Mehta, 2003; Jennison & Turnbull, 2006*.
- From J & T: "the flexibility of unplanned adaptive designs comes at a price" ... "standard error-spending tests provide efficient designs, but it is still possible to fall back on flexible methods".

Proposed ~~solution~~ strategy

Magirr, Stallard & Jaki, 2013

Unfortunately,

FWER control + flexibility + sufficient statistics = impossible

One can, however,

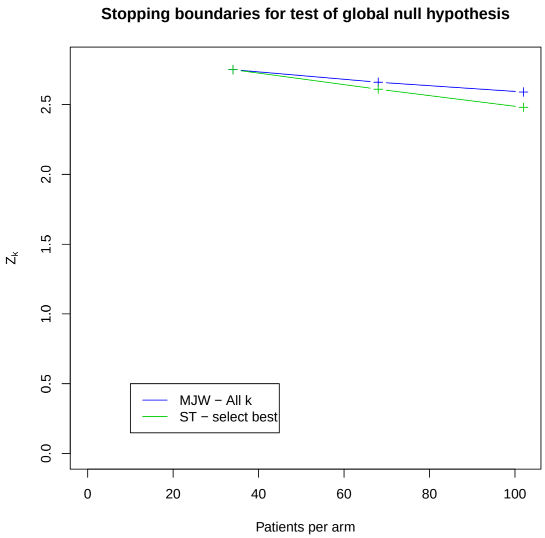
1. Set up the trial using a group-sequential multi-arm design (either ST or MJW).
2. If the unexpected happens → calculate the **conditional error**.
3. Adjust the stopping boundaries to take account of unplanned design changes.

Example ($\approx$ re-design of Galantamine trial)

- Consider a 3-stage trial comparing 3 experimental treatments with control.
- Suppose

$$\alpha_1^* = (1/3)0.025, \qquad \alpha_2^* = (2/3)0.025, \qquad \alpha_3^* = 0.025.$$

- Also, suppose $l_1 = l_2 = -\infty$ (non-binding futility boundary).
- Sample size is $n = 34$ patients per arm per stage.
- This ensures LFC power of 0.8 given $\delta = 0.5\sigma$ and $\delta_0 = 0.2\sigma$.


Stopping boundaries for test of global null hypothesis
Z_k
0.0
0.5
1.0
1.5
2.0
2.5
0
20
40
60
80
100
Patients per arm
MJW − All k
ST − select best

First interim analysis

- Suppose $Z_1^{(1)} = 2$, $Z_2^{(1)} = 1.1$ and $Z_3^{(1)} = 1$.
- None of the test statistics cross the stopping boundary.
- However, treatment 1 is dropped from the study due to safety concerns.

Conditional error (MJW design)

1. Find

$$\alpha_2^*(X_1) = P_0 \bigcup_{k=1}^{3} \{Z_k > u_2\} \mid X_1$$

and

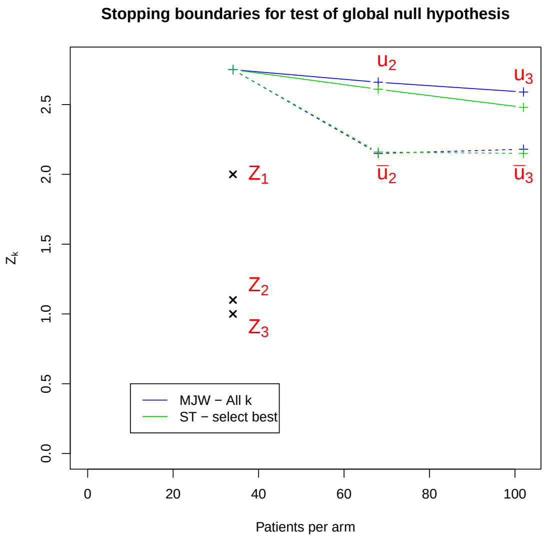
$$\alpha_3^*(X_1) = P_0 \bigcup_{k=1}^{3} \{Z_k \text{ crosses boundary}\} \mid X_1,$$

where X_1 is the first-stage data.

2. Find adjusted stopping boundary, $\bar{u}_2, \bar{u}_3$, such that

$$P_0 \bigcup_{k=2}^{3} \{Z_k > \bar{u}_2\} \mid X_1 = \alpha_2^*(X_1)$$

$$P_0 \bigcup_{k=2}^{3} \{Z_k \text{ crosses (adjusted) boundary}\} \mid X_1 = \alpha_3^*(X_1)$$


Stopping boundaries for test of global null hypothesis
u2
u3
ū2
ū3
Z1
Z2
Z3
MJW − All k
ST − select best
Zk
0.0
0.5
1.0
1.5
2.0
2.5
0
20
40
60
80
100
Patients per arm

Comments

- Effect of (unplanned) dropping of treatment arm $\rightarrow$ boundary lowered $\rightarrow$ more power for remaining hypotheses.
- Additional complexity: To control the FWER here, one must apply the closure principle, i.e., one must consider all $2^3 - 1$ intersection null hypothesis $\rightarrow$ each intersection null hypothesis requires its own adjusted boundaries.
- All calculations involve well-known properties of the multivariate normal distribution.
- See Magirr, Stallard & Jaki (2013, submitted) for details.

A small simulation study

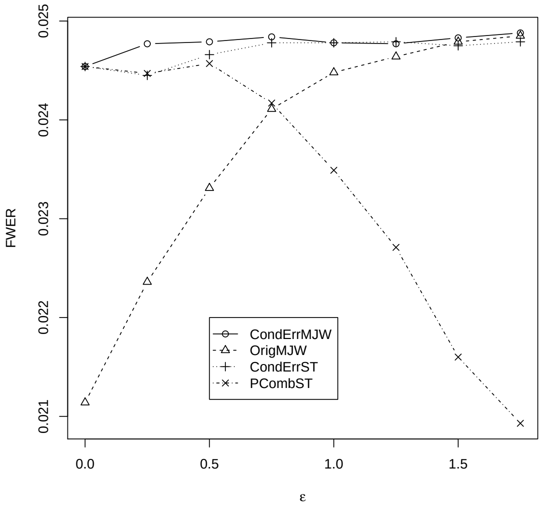
1. How large is the power gain due to modified upper stopping boundary?
2. How much power is lost from using non-sufficient statistics?

- Suppose we distort the pre-specified selection rule by, at the jth interim analysis, selecting

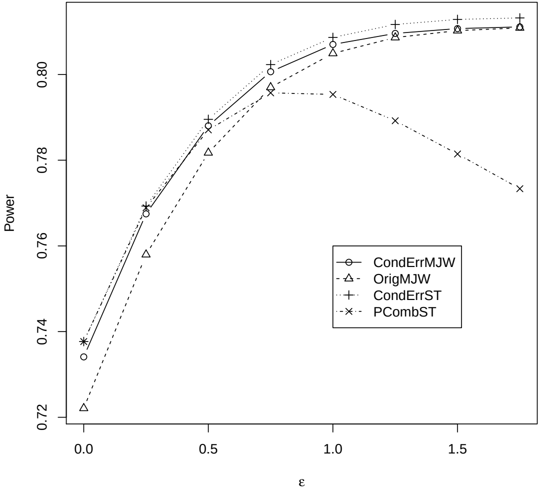
$$T^{(j+1)} = \left\{ k : Z_k^{(j)} \geq max_{k' \in T^{(j)}} Z_{k'}^{(j)} - \epsilon \right\}.$$

- I.e., "continue with all treatments within ϵ of the best".
- Simulating the trial 100,000 times...

Simulation study - FWER



Simulation study - Power



Conclusions

- Flexibility can be added to multi-arm group sequential studies.
- It is important to put as much effort as possible into finding an appropriate multi-arm group sequential design $\rightarrow$ one should only break the sufficiency principle if absolutely necessary (something totally unexpected happens).
- In principle, same techniques could be used to change the sample size or add additional interim analyses.
- The computation only involves MVN probabilities (but very fiddly). I will put all the R code into our "MAMS" package.

References

- Dunnett, C.W. (1984) Selection of the best treatment in comparison to a control with an application to a medical trial. In *Design of Experiments: Ranking and Selection*
- Magirr, D., et al. (2012) A generalized Dunnett test for multi-arm, multi-stage clinical studies with treatment selection. *Biometrika* **99**, 494-501.
- Follmann, D.A., et al. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* **50**, 325-336.
- Stallard, N. & Todd, S. (2003) Sequential designs for phase III clinical trials incorporating treatment selection. *Stats in Med* **22** 686-703.
- Jennison, C. and Turnbull, B.W. (2006) Adaptive and nonadaptive group sequential tests. *Biometrika* **93** 1-21.
- Bauer, P. & Köhne, K. (1994) Evaluation of experiments with adaptive interim analyses. *Biometrics* **50** 1029-1041.
- Proschan, M. & Hunsberger, S. (1995) Designed extension of studies based on conditional power. *Biometrics* **51** 1315-1324.

References

- Marcus, R., et al. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655-660.
- Burman, C-F. & Sonesson, C. Are flexible designs sound? *Biometrics* **62** 664-683.
- Posch, M., et al. (2005). Testing and Estimation in Flexible Group Sequential Designs with Adaptive Treatment Selection. *Statistics in Medicine* **24**, 3697-3714.
- Bauer, P. and Kieser, M. (1999) Combining Different Phases in the Development of Medical Treatments Within a Single Trial. *Statistics in Medicine* **18**, 1833-1848.
- Koenig, F., et al. (2008) Adaptive Dunnett Tests for Treatment Selection. *Statistics in Medicine* **27**, 1612-1625.
- Magirr, D., Stallard, N. & Jaki, T. (2013) Flexible sequential designs for multi-arm clinical trials. *Submitted*.