Designing a Multi-test Multi-stage Clinical Trial

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Workshop on: Advanced Statistical Designs to Empower Biomarker-driven Clinical Trials

Bath

April, 2024

1. Examples

Multi-stage clinical trials that test multiple hypotheses include:

- Seamless Phase 2/3 trials
- Trials comparing several experimental treatments to a control
  - Two versions of a new treatment vs control (GATSBY)
  - Multiple new treatments vs control (Umbrella trials)
- Enrichment designs
  - Two patient subgroups
  - Ordered subgroups based on a biomarker value

Aim: To reach conclusions, rejecting zero, one or more null hypotheses while controlling the **overall** type I error rate.

2. Controlling the Familywise Error Rate

Suppose we have h null hypotheses, H_i: $\theta_i \leq 0$ for $i = 1, \ldots, h$.

A procedure's **familywise error rate** when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h)$ is

$$P_{\boldsymbol{\theta}}\{\text{Reject } H_i \text{ for some } i \text{ with } \theta_i \leq 0\}.$$

The familywise error rate is controlled **strongly** at level α if this error rate is at most α for all possible combinations of θ_i values.

Then

$$P_{\boldsymbol{\theta}}\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_h).$$

Using such a procedure, the probability of choosing to focus on a parameter θ_{i^*} and then falsely claiming significance for the associated null hypothesis H_{i^*} is at most α.

Closed testing procedures (Marcus et al. *Biometrika*, 1976)

Suppose we have null hypotheses H_i, $i = 1, \ldots, h$.

For each subset I of $\{1, \ldots, h\}$, define the intersection hypothesis

$$H_I = \cap_{i \in I} H_i.$$

Construct a "local" level α test of each intersection hypothesis H_I, i.e., a test which rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed testing procedure

The simple hypothesis H_j: $\theta_j \leq 0$ is rejected overall if, and only if, the "local" tests reject H_I for every set I containing index j.

Closed testing procedures

Proof of strong control of familywise error rate

Let $\tilde{I}$ be the set of indices of all true hypotheses H_i.

Since $H_{\tilde{I}}$ is true, $P\{\text{Reject } H_{\tilde{I}}\} = \alpha$.

For a familywise error to be committed, $H_{\tilde{I}}$ must be rejected.

Hence, the probability of a familywise error is no greater than α.

Testing an intersection hypothesis

Suppose the intersection hypothesis $H_I = \cap_{i \in I} H_i$ is the intersection of m simple hypotheses.

For each $i \in I$, let P_i be the 1-sided P-value for testing H_i.

Denote the ordered values of the P_i by $P_{[1]} \leq P_{[2]} \leq \ldots \leq P_{[m]}$.

Bonferroni adjustment

The overall P-value for testing H_I is defined to be $P_I = m\, P_{[1]}$.

Simes' method (Biometrika, 1986)

The Simes P-value for H_I is

$$P_I = \min_{k=1,\ldots,m} (m\, P_{[k]}/k).$$

The Simes method is valid when the P_i are independent or positively dependent. It is less conservative than Bonferroni.

Dunnett's method (JASA, 1955)

Suppose m treatments are compared with a control, and responses are normal with known variance.

Each null hypothesis H_i says treatment i is no better than control.

We are to test the intersection hypothesis $H_I = \cap_{i \in I} H_i$.

Denote the Z-statistic arising from the test of H_i by Z_i.

The Z_i have a multivariate normal distribution with a known covariance matrix.

The P-value for testing H_I using Dunnett's test is

$$P\{\max_{i \in I} Z_i > z^*\},$$

where z^* is the observed value of $\max_{i \in I} Z_i$, and the probability is calculated assuming each $E(Z_i) = 0$.

3. Multi-stage adaptive designs: Combination tests

Suppose we run a clinical trial adaptively in two stages:

Set the design of Stage 1, then conduct this part of the trial,

Analyse results from Stage 1,

Consider external information, if appropriate.

Set the design of Stage 2, informed by Stage 1 results and external information,

Conduct Stage 2,

Analyse the results from Stage 2.

How can we test a null hypothesis with proper protection of the type I error rate?

Combination tests (Bauer & Köhne, *Biometrics*, 1994)

Let θ denote the treatment effect *vs* control for a specified form of the treatment, patient population and endpoint.

We test H_0: $\theta \leq 0$ against $\theta > 0$, with type I error rate α at $\theta = 0$.

Define one-sided P-values $P^{(1)}$ and $P^{(2)}$ from hypothesis tests of H_0 based on Stage 1 and Stage 2 data, respectively.

Under $\theta = 0$

$P^{(1)} \sim U(0,1)$.

Conditional on Stage 1 data and Stage 2 design, $P^{(2)} \sim U(0,1)$.

Thus, $P^{(1)}$ and $P^{(2)}$ are independent $U(0,1)$ variates when $\theta = 0$.

Hence $P^{(1)}$ and $P^{(2)}$ can be combined in an overall test of H_0.

The inverse normal combination test

Initial design

Specify use of the **inverse normal test** for hypothesis H_0, with weights w_1 and w_2 where $w_1^2 + w_2^2 = 1$.

Design Stage 1, fixing sample size and test statistic.

Stage 1

Observe the one-sided P-value, $P^{(1)}$, based on Stage 1 data.

Compute $Z^{(1)} = \Phi^{-1}(1 - P^{(1)})$.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, $P^{(2)}$, based **only** on Stage 2 data.

Compute $Z^{(2)} = \Phi^{-1}(1 - P^{(2)})$.

NB Under $\theta = 0$, $Z^{(1)} \sim N(0,1)$, $Z^{(2)} \sim N(0,1)$, independent.

The inverse normal combination test

The combination test is based on the statistic $w_1 Z^{(1)} + w_2 Z^{(2)}$.

Under $\theta = 0$, $Z^{(1)}$ and $Z^{(2)}$ are independent $N(0,1)$ so, with $w_1^2 + w_2^2 = 1$,

$$w_1 Z^{(1)} + w_2 Z^{(2)} \sim N(0,1).$$

Hence, for an overall one-sided test with type I error rate α, we reject H_0 if

$$w_1 Z^{(1)} + w_2 Z^{(2)} > \Phi^{-1}(1-\alpha).$$

If $\theta < 0$, then $Z^{(1)}$ and $Z^{(2)}$ are stochastically smaller than $N(0,1)$ random variables and the type I error rate is less than α.

A K-stage combination test (Lehmacher & Wassmer, *Biometrics*, 1999)

Start by defining a standard form of group sequential test

At analysis k, statistics Z_k (based on the cumulative data) are compared with critical values a_k and b_k.

If $Z_k < a_k$ or $Z_k > b_k$ the test stops, rejecting H_0 if $Z_k > b_k$.

Values of a_k and b_k are set so the type I error probability is α.

For $k = 1, \ldots, K$, let $Z^{(k)}$ be the standardised test statistic based on Stage k data alone, and write the cumulative Z-statistics as

$$Z_k = (w_1 Z^{(1)} + \ldots + w_k Z^{(k)}) / (w_1^2 + \ldots + w_k^2)^{1/2}. \quad (1)$$

In the adaptive trial design, calculate each $Z^{(k)}$ based on Stage k data alone and substitute these values into (1).

Applying the stopping rule with critical values a_k and b_k gives a group sequential test with type I error rate α.

A Closed Testing Procedure with Combination Tests

It is straightforward to put together these two ingredients of a design.

The Closed Testing Procedure requires a "local" level α test of each individual hypothesis or intersection hypothesis.

Each level α test is created by combining the P-values or Z-values based on data accrued in successive stages of the study. (In the case of an intersection hypothesis, this may be a Bonferroni, Simes or Dunnett P-value or Z-value.)

At each stage of group sequential testing a hypothesis H_i can be rejected overall if every intersection hypothesis H_I with $i \in I$ has been rejected by its local, level α test.

4. Example: A Multi-arm Multi-stage (MAMS) trial

Suppose 3 treatments, low, medium and high doses of a new drug, are to be compared against a control in a 4-stage trial.

We specify:

A Closed Testing Procedure,

Dunnett's method to be used to create stage-wise Z-values for intersection hypotheses,

Lehmacher-Wassmer, 4-stage combination tests for each H_I based on ρ-family error spending tests (see JT, 2000) with

$\rho = 2$, $\alpha = 0.025$, no futility boundary (so each $a_k = -\infty$).

The null hypothesis H_j: $\theta_j \leq 0$ can be rejected globally if the Lehmacher-Wassmer "local" tests reject each H_I with $j \in I$.

Each treatment may be discontinued at any point for positive or negative reasons.

Example: Stage 1 data

Suppose the first stage produces Z-statistics $Z_1^{(1)}$, $Z_2^{(1)}$, and $Z_3^{(1)}$ for the three treatments, as shown below.

Treatment j	$Z_j^{(1)}$
1	1.26
2	1.84
3	2.76

We shall apply Dunnett's rule to find the Z-value $Z_I^{(1)}$ for each intersection hypothesis H_I.

At Stage 1, the $Z_I^{(1)}$ are also the cumulative Z-values, $Z_{I,1}$, that appear in the Lehmacher-Wassmer test.

The Lehmacher-Wassmer testing boundary has

$$b_1 = 2.96, \quad b_2 = 2.56, \quad b_3 = 2.30, \quad b_4 = 2.09,$$

so we need to see $Z_{I,1} = Z_I^{(1)} \geq 2.96$ to reject H_I at this stage.

Example: Stage 1 data

Applying Dunnett's rule, Z-values for intersection hypotheses are

Hypothesis H_I	$Z_I^{(1)}$
$H_{\{1\}}$	1.26
$H_{\{2\}}$	1.84
$H_{\{3\}}$	2.76
$H_{\{1,2\}}$	1.56
$H_{\{1,3\}}$	2.54
$H_{\{2,3\}}$	2.54
$H_{\{1,2,3\}}$	2.41

As already noted, the $Z_I^{(1)}$ are also the cumulative Z-values, $Z_{I,1}$, that appear in the Lehmacher-Wassmer test.

As each $Z_{I,1} = Z_I^{(1)} < b_1 = 2.96$, no hypotheses are rejected here.

We suppose the trial continues with all 3 treatments still active.

Example: Stage 2 data

Results in Stage 2 (only) produce the Z-statistics $Z_1^{(2)}$, $Z_2^{(2)}$ and $Z_3^{(2)}$ shown below.

Treatment j	$Z_j^{(2)}$
1	-0.45
2	2.21
3	0.71

From these, we compute the Dunnett Z-values, $Z_I^{(2)}$, for each intersection hypothesis H_I.

Then, to apply the Lehmacher-Wassmer test, we calculate the cumulative Z-value for each H_I

$$Z_{I,2} = \frac{Z_I^{(1)} + Z_I^{(2)}}{\sqrt{2}}.$$

Example: Results after Stages 1 and 2

H_I	$Z_I^{(1)} = Z_{I,1}$	$Z_I^{(2)}$	$Z_{I,2}$
$H_{\{1\}}$	1.26	-0.45	0.57
$H_{\{2\}}$	1.84	2.21	2.86
$H_{\{3\}}$	2.76	0.71	2.45
$H_{\{1,2\}}$	1.56	1.96	2.49
$H_{\{1,3\}}$	2.54	0.34	2.04
$H_{\{2,3\}}$	2.54	1.96	3.18
$H_{\{1,2,3\}}$	2.41	1.81	2.98

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{2\}}$, $H_{\{2,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{I,2} > b_2 = 2.56$.

However, $H_{\{1,2\}}$ is not rejected so the Closed Testing Procedure does not allow global rejection of H_2.

Suppose the high dose Treatment 3 is dropped for safety reasons, so the trial continues with Treatments 1 and 2 and the control.

Example: Stage 3 data

Results in Stage 3 (only) produce Z-statistics $Z_1^{(3)}$ and $Z_2^{(3)}$

Treatment j	$Z_j^{(3)}$
1	0.90
2	1.41
3	—

In computing the Dunnett Z-value for an intersection hypothesis H_I with $3 \in I$, we set $Z_I^{(3)}$ equal to $Z_{I'}^{(3)}$ where $I' = I \setminus \{3\}$.

This cannot be done for $I = \{3\}$ — but that is not a problem as we are no longer interested in the global test of H_3.

The cumulative Z-value for each H_I after the first 3 stages is

$$Z_{I,3} = \frac{Z_I^{(1)} + Z_I^{(2)} + Z_I^{(3)}}{\sqrt{3}}.$$

Example: Results after Stages 1, 2 and 3

H_I	$Z_I^{(1)} = Z_{I,1}$	$Z_I^{(2)}$	$Z_{I,2}$	$Z_I^{(3)}$	$Z_{I,3}$
$H_{\{1\}}$	1.26	-0.45	0.57	0.90	0.99
$H_{\{2\}}$	1.84	2.21	2.86	1.41	3.15
$H_{\{3\}}$	2.76	0.71	2.45	—	—
$H_{\{1,2\}}$	1.56	1.96	2.49	1.10	2.67
$H_{\{1,3\}}$	2.54	0.34	2.04	0.90	2.19
$H_{\{2,3\}}$	2.54	1.96	3.18	1.41	3.41
$H_{\{1,2,3\}}$	2.41	1.81	2.98	1.10	3.07

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{2\}}$, $H_{\{1,2\}}$, $H_{\{2,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{I,3} > b_3 = 2.30$.

Thus, H_2 can be rejected globally and Treatment 2 declared superior to the control.

Suppose Treatment 2 is discontinued at this point and the trial continues with the low dose Treatment 1 and the control.

Example: Stage 4 data

Results in Stage 4 (only) produce the single Z-statistic $Z_1^{(4)}$.

Treatment j	$Z_j^{(4)}$
1	2.07
2	—
3	—

We shall use $Z_1^{(4)}$ to create Z-statistics for intersection hypotheses H_I involving Treatment 1.

We can then conduct the final analysis of the Lehmacher-Wassmer tests of these hypotheses using the test statistics

$$Z_{I,4} = \frac{Z_I^{(1)} + Z_I^{(2)} + Z_I^{(3)} + Z_I^{(4)}}{2}.$$

Example: Results after Stages 1, 2, 3 and 4

H_I	$Z_I^{(1)}$	$Z_I^{(2)}$	$Z_{I,2}$	$Z_I^{(3)}$	$Z_{I,3}$	$Z_I^{(4)}$	$Z_{I,4}$
$H_{\{1\}}$	1.26	-0.45	0.57	0.90	0.99	2.07	1.89
$H_{\{2\}}$	1.84	2.21	2.86	1.41	3.15	—	—
$H_{\{3\}}$	2.76	0.71	2.45	—	—	—	—
$H_{\{1,2\}}$	1.56	1.96	2.49	1.10	2.67	2.07	3.35
$H_{\{1,3\}}$	2.54	0.34	2.04	0.90	2.19	2.07	2.93
$H_{\{2,3\}}$	2.54	1.96	3.18	1.41	3.41	—	—
$H_{\{1,2,3\}}$	2.41	1.81	2.98	1.10	3.07	2.07	3.69

The Lehmacher-Wassmer tests reject intersection hypotheses $H_{\{1,2\}}$, $H_{\{1,3\}}$ and $H_{\{1,2,3\}}$ since they have $Z_{I,4} > b_4 = 2.09$.

However, $H_{\{1\}}$ is not rejected, so the Closed Testing Procedure does not allow global rejection of H_1.

Example: Conclusions

The conclusions from the study are that:

The medium dose, Treatment 2, was shown to be superior to the control in a testing procedure with familywise type I error rate $\alpha = 0.025$.

The low dose, Treatment 1, was not found to be superior to the control.

The high dose, Treatment 3, was found to have safety problems and was dropped half way through the study.

5. Creating efficient designs

We have shown how to construct designs that protect the familywise error rate.

We can also choose elements of a design to "optimise" aspects of its performance.

In a sequential design, we can aim to reduce average sample size subject to controlling type I error and achieving a specified power.

In optimising a design that tests multiple hypotheses, we need to define an overall measure of the value of a final set of outcomes, depending on which null hypotheses are rejected.

Then, we can try to maximise the expected value of this "gain function", either under a certain set of parameter values or integrated over a prior distribution.

Creating efficient designs

Optimising every element of a complex design may not be feasible.

However, one can fix certain elements to control familywise error and then optimise the remaining parts of the design.

The process of assessing the value of a final set of outcomes and identifying the parameter values under which to optimise performance can be an instructive exercise in itself.

Some references:

Barber, S. and Jennison, C., 2002. Optimal asymmetric one-sided group sequential tests. *Biometrika*, **89**, 49–60.

Hampson, L.V. and Jennison, C., 2015. Optimizing the data combination rule for seamless phase II/III clinical trials. *Statist. in Med.*, **34**, 39–58.

Burnett, T. and Jennison, C., 2021. Adaptive enrichment trials: What are the benefits? *Statist. in Med.*, **40**, 690–711.