Multiple Testing in Clinical Trials

February 29, 2020

Contents

What is the issue?

When performing MANY (e.g. independent) tests, we expect to have at least one significant result by chance even though no difference exists.

Probability of at least one false significant result

Number of tests	Probability
1	0.05
2	0.0975
5	0.226
10	0.401
50	0.923

P(at least one false positive result)
= 1 - P(zero false positive results)
= $1 - (1 - .05)^k$

The multiplicity problem

- Testing a family (many) of hypotheses risk of giving us significant results just by chance.
- We want to find methods "Multiple Testing procedures" (MTP) to control this global risk (family wise error rate).
- The same problem arises when considering many confidence intervals simultaneously.

Family-wise error rates

- A family is a collection of a priori stated null hypotheses

$$F = \{H_1, \ldots, H_n\}.$$

- Test statistics $t_1, t_2, \ldots, t_n$.
- p-values: $p_1, p_2, \ldots, p_n$.
- MTPs are commonly designed to control the Type I Familywise Error Rate (FWER):

$$\text{FWER} = P\{\text{Reject at least one true } H_i\} \leq \alpha$$

Issues of multiplicity in clinical trials

- Multiple endpoints (efficacy and safety)
- Multiple treatment arms or doses of a drug
- Interim analyses (group sequential trials)
- Subgroup analyses
- Data-snooping or data-fishing
- Chance of false positives increases if no adjustment for multiplicity is made.
- Multiple test procedures (MTPs) control frequency of false positives.

Regulatory requirements

EMEA/CPMP's (2002) Points to Consider on Multiplicity Issues …:

(*from Section 2.5*)

As a general rule it can be stated that control of the family-wise type-I error in the strong sense (i.e. application of closed test procedures) is a minimal prerequisite for confirmatory claims.

(*from Section 7*)

It is therefore necessary that the statistical procedures planned to deal with, or to avoid, multiplicity are fully detailed in the study protocol or in the statistical analysis plan to allow an assessment of their suitability and appropriateness.

Additional claims on statistical significant and clinically relevant findings based on secondary variables or on subgroups are possible only after the primary objective of the clinical trial has been achieved, and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy

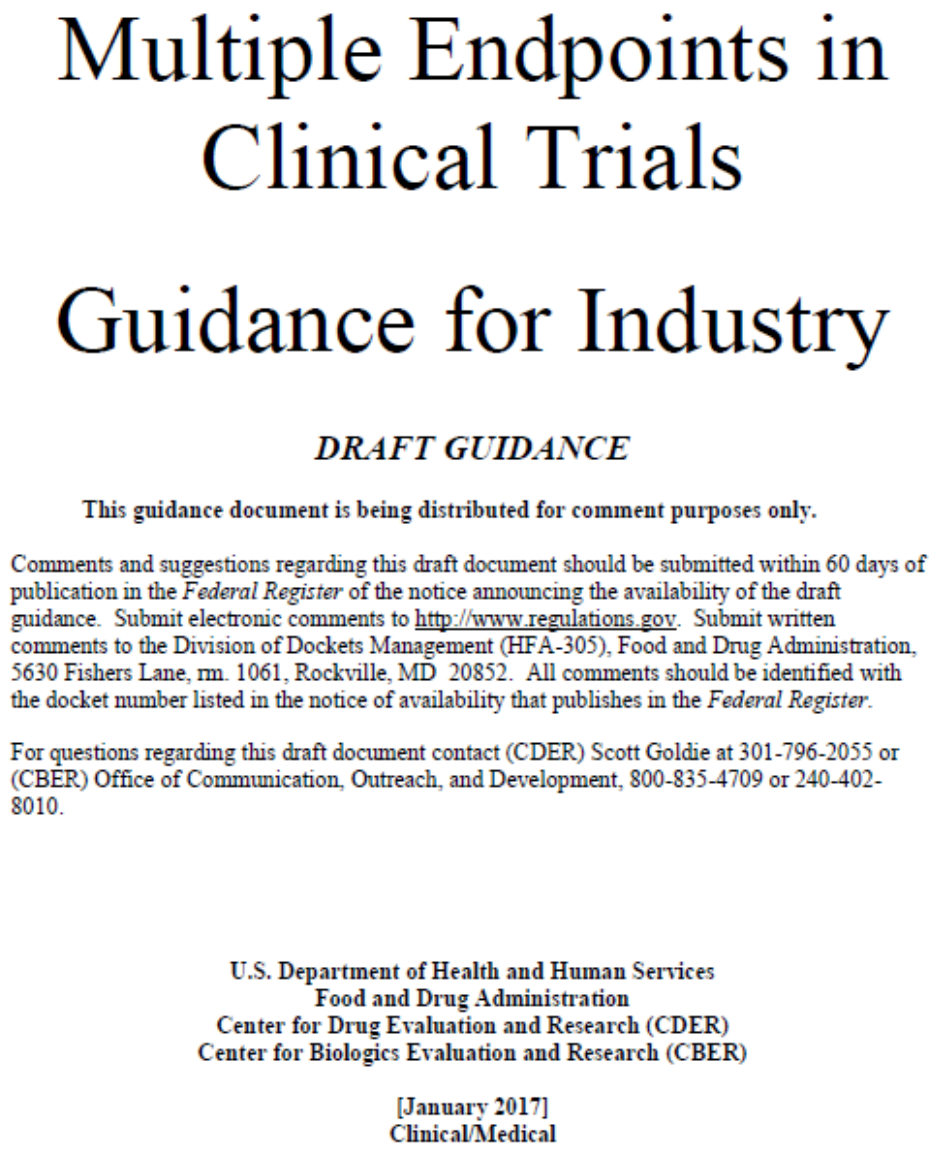
Simple strategies

A. Multiple treatments

- Arrange the treatment comparisons in order of importance
- Decide which comparisons should belong to the confirmatory analysis
- Decide a way to control the error of false significances for these comparisons

Simple strategies
B. Multiple endpoints

- Find out which variables are needed to answer the primary objective of the study
- Look for possibilities to combine the variables, e.g. composite endpoints, global measures (QoL, index etc.)
- Decide a way to control the error of false significances for these variables



Multiple Endpoints in Clinical Trials

Guidance for Industry

DRAFT GUIDANCE

This guidance document is being distributed for comment purposes only.

Comments and suggestions regarding this draft document should be submitted within 60 days of publication in the *Federal Register* of the notice announcing the availability of the draft guidance. Submit electronic comments to http://www.regulations.gov. Submit written comments to the Division of Dockets Management (HFA-305), Food and Drug Administration, 5630 Fishers Lane, rm. 1061, Rockville, MD 20852. All comments should be identified with the docket number listed in the notice of availability that publishes in the *Federal Register*.

For questions regarding this draft document contact (CDER) Scott Goldie at 301-796-2055 or (CBER) Office of Communication, Outreach, and Development, 800-835-4709 or 240-402-8010.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)

[January 2017]
Clinical/Medical

Example

- Trial to evaluate the effects of lisinopril on mortality and morbidity of patients with heart disease (similar to Packer et al. studies (1996. 1999) on amlodipine and lisinopril).
- Two co-primary endpoints:
  - All-cause mortality
  - All-cause mortality or all-cause hospitalization
- Win criterion: Win on *at least one* endpoint (classical multiple comparisons problem)

Example

- Trial to evaluate the effects of donepezil on cognition and global changes in patients with mild to moderate Alzheimer's disease.
- Two co-primary endpoints:
  - Alzheimer's disease assessment scale-Cognitive subscale (ADAS-Cog)
  - Clinician global impression change (CGIC)
- Win criterion: Win on *both* endpoints

Simple strategies
C. Multiple time points

- Find out which time points are the most relevant for the treatment comparison
- If no single time point is most important, look for possibilities to combine the time points, e.g. chnge from baseline, average over time (AUC etc.)
- Decide a way to control the error of false significances if more than one important time point

Simple strategies

D. Interim analyses

- Decide if the study should stop for safety and/or efficacy reasons
- Decide the number of interim analyses
- To control the error of a false significance (stopping the study), decide how to spend the total significance level on the interim and final analysis

Simple strategies

E. Subgroup analyses

- Subgroup analyses are usually not part of a confirmatory analysis
- Restrict the number of subgroup analyses
- Use only subgroups of sufficient size
- All post-hoc subgroup analyses are considered exploratory

What's multiplicity got to do with me?

- "I (am a Bayesian so I) do not agree with the principles behind adjustment"
  - OK, but regulatory authorities will (may) take a different view

- "I work in oncology where we generally use all patients, have 1 treatment comparison, 1 primary endpoint (Time to event) and a small number of secondary endpoints"

  - Still multiplicity issues around secondary endpoints
  - Not always this simple:
    - 2 populations e.g. all, biomarker positive group
    - More than 1 treatment comparison e.g. experimental vs. control, experimental + control vs. control

What's multiplicity got to do with me?

- "I work in early phase trials"
  - Phase II used for internal decision making so we do not have to take account of the multiplicity (trials would become too big if we did)

- Agree, but issues of multiplicity still apply
  - We need to understand any increase in the risk of a false positive finding and as long as this is understood it may be acceptable

Methods based on p-values

- Correlations between endpoints are unknown, so parametric procedures based on multivariate test statistics can't be exactly used.
- Marginal p-values are readily available (but ignore correlations).
- Marginal p-values may come from diverse tests, e.g., t-tests, χ^2-tests, logrank tests, etc.
    - Holm Procedure
    - Hochberg Procedure

Bonferroni

- N different null hypotheses H_1, … H_N
- Calculate corresponding p-values p_1, … p_N
- Reject H_k if and only if $p_k < \alpha/N$

Variation: The limits may be unequal as long as they sum up to α

Conservative

- $P(A_i)$ = P(reject H_{0i} when it is true) $\leq \frac{\alpha}{N}$

$$P\left(\bigcup_{i=1}^{N} A_i\right) \leq \sum_{i=1}^{N} P(A_i) \leq \sum_{i=1}^{N} \frac{\alpha}{N} = N\frac{\alpha}{N} = \alpha$$

Reject at least one hypthesis falsely

Holm

- Holm (1979)
- Step-down algorithm

$$\begin{array}{ccccccc} H_{(1)} & & H_{(2)} & & & & H_{(n)} \\ p_{(1)} & \leq & p_{(2)} & \leq & \cdots & \leq & p_{(n)} \\ \frac{\alpha}{n} & & \frac{\alpha}{n-1} & & & & \frac{\alpha}{1} \end{array}$$

- Begin testing with $p_{(1)}$ & continue as long as you get rejections. If at the ith step $p_{(i)} > \frac{\alpha}{n-i+1}$ then accept $H_{(i)}$ and all the remaining hypotheses.

Hochberg

- Hochberg (1988):

- Step-up algorithm

$$\begin{array}{ccccccc} H_{(1)} & & H_{(2)} & & & & H_{(n)} \\ p_{(1)} & \leq & p_{(2)} & \leq & \cdots & \leq & p_{(n)} \\ \frac{\alpha}{n} & & \frac{\alpha}{n-1} & & & & \frac{\alpha}{1} \end{array}$$

- Begin testing with $p_{(n)}$ & continue as long as you get acceptances. If at the ith step $p_{(i)} < \frac{\alpha}{n-i+1}$ then reject $H_{(i)}$ and all the remaining hypotheses.

A simple example

- Assume we performed N=5 tests of hypothesis simultaneously and want the result to be at the level 0.05.
- The p-values obtained are as in the table
- The p values might come from different tests and the test statistics might be correlated or not

p(1)	**0.009**
p(2)	**0.011**
p(3)	**0.012**
p(4)	**0.134**
p(5)	**0.512**

- **Bonferroni:**
  - 0.05/5=0.01. Since only p(1) is less than 0.01 we reject H(1) but accept the remaining hypotheses.

p(1)	0.009
p(2)	0.011
p(3)	0.012
p(4)	0.134
p(5)	0.512

$p_i < \alpha/N$

- **Holm:**
  - p(1), p(2) and p(3) are less than 0.05/5, 0.05/4 and 0.05/3 respectively so we reject the corresponding hypotheses H(1), H(2) and H(3). But p(4) = 0.134 > 0.05/2=0.025 so we stop and accept H(4) and H(5).

p(1)	0.009
p(2)	0.011
p(3)	0.012
p(4)	0.134
p(5)	0.512

$p_{(i)} > \frac{\alpha}{n-i+1}$

- **Hochberg:**
  - 0.512 is not less than 0.05 so we accept H(5)
  - 0.134 is not less than 0.025 so we accept H(4)
  - 0.012 is less than 0.0153 so we reject H(1),H(2) and H(3)

p(1)	0.009
p(2)	0.011
p(3)	0.012
p(4)	0.134
p(5)	0.512

$p_{(i)} < \frac{\alpha}{n-i+1}$

Ordered Hypotheses: Fixed sequence

- In some problems hypotheses are a priori ordered based on importance, e.g., ordered doses.

$$H_1 \to H_2 \to \cdots \to H_n.$$

- Fixed sequence procedure: Starting with H_1, reject each H_i if $p_i \leq \alpha$. Continue testing as long as rejections occur. Stop testing and accept all the remaining hypotheses if an acceptance occurs.
- No α-adjustment (Maurer, Hothorn & Lehmacher 1995).

(Example: Dose finding)

Summary

- In terms of power, Hochberg $>$ Holm $>$ Bonferroni.
- Hochberg requires p-values to be independent or positively correlated; no such restriction on Holm and Bonferroni.
- Fixed sequence could be used for a priori ordered hypotheses.

Example

Placebo: $n = 1596$, Treatment: $n = 1568$, $\alpha = 0.025$

Endpoint	Event Rate (%)		z-statistic	1-sided p-value
	Placebo	Treatment		
E1	44.8	41.1	2.102	0.018
E2	83.8	80.8	2.211	0.014

Both significant?

- Bonferroni Procedure: Both $p_1 = 0.018$ and $p_2 = 0.014 > \alpha/2 = 0.0125$, so declare both not significant.
- Holm Procedure: $p_{(1)} = p_2 = 0.014 > 0.0125$, so stop testing and declare both not significant.
- Hochberg Procedure: $p_{(2)} = p_1 = 0.018 < \alpha = 0.025$, so stop testing and declare both significant.
- Fixed Sequence Procedure: $p_1 = 0.018 < \alpha = 0.025$ and $p_2 = 0.014 < \alpha = 0.025$, so declare both significant.

Methods for constructing multiple testing procedures

- 4.1 Union-Intersection (At Least One) Method
- 4.2 Intersection-Union (All or None) Method
- 4.3 Closure Method

We only definet for the closure method

Closed testing procedures

- To use this procedure, start with the global test $H=\bigcap_{i=1}^{n} H_i$
- If this test is rejected at level α, proceed each subset of (n-1) hypotheses.
- As long as hypotheses continue to be rejected at level α, contiue testing
- Eventually wou will reach subsets of size 1, i.e. the individual hypotheses H_i
- Such procedures control the familiwise error rate i.e. all tests are tested at level α
- Holm's method is a special case of a closed test procedure

An example

- Suppose there are three hypotheses H_1,H_2, and H_3 to be tested and the overall type I error rate is 0.05. Then H_1 can be rejected at level α if
  - $H_1 \cap H_2 \cap H_3$,
  - $H_1 \cap H_2$, $H_1 \cap H_3$
  - H_1
- can all be rejected using valid tests with level 0.05.

Adjusted p-values

- The adjusted *p*-values for Holm–Bonferroni method are:
- **Example.** Suppose we have ten p-values as in the table below. Most of them are impressively small, even after accounting for the fact that we have 10 of them.

				$p(i) < \frac{\alpha}{K-i+1}$ Holm adjustment	$p_i < \alpha/K$ Bonferroni adjustment
i	p(i)	(K-i+1)	(K-i+1)p(i)	max(p*(1),p*(2),... ,p*(i))	10*p(i)
1	0.0002	10	0.0020	0.0020	0.0020
2	0.0011	9	0.0099	0.0099	0.0110
3	0.0012	8	0.0096	0.0099	0.0120
4	0.0015	7	0.0105	0.0105	0.0150
5	0.0022	6	0.0132	0.0132	0.0220
6	0.0091	5	0.0455	0.0455	0.0910
7	0.0131	4	0.0524	0.0524	0.1310
8	0.0152	3	0.0456	0.0524	0.1520
9	0.0311	2	0.0622	0.0622	0.3110
10	0.1986	1	0.1986	0.1986	1.0000

Adjusted p-values

Section 6: p-value adjustment for multiple comparisons

For studies with multiple outcomes, p-values can be adjusted to account for the multiple comparisons issue. The '`p.adjust( )`' command in R calculates adjusted p-values from a set of un-adjusted p-values, using a number of adjustment procedures.

Adjustment procedures that give strong control of the family-wise error rate are the Bonferroni, Holm, Hochberg, and Hommel procedures.

Adjustments that control for the false discovery rate, which is the expected proportion of false discoveries among the rejected hypotheses, are the Benjamini and Hochberg, and Benjamini, Hochberg, and Yekutieli procedures.

To calculate adjusted p-values, first save a vector of un-adjusted p-values. The following example is from a study comparing two groups on 10 outcomes through t-tests and chi-square tests, where 3 of the outcomes gave un-adjusted p-values below the conventional 0.05 level. The following calculates adjusted p-values using the Bonferroni, Hochberg, and Benjamini and Hochberg (BH) methods:

```
> pvalues <- c(.002, .005, .015, .113, .222, .227, .454, .552, .663, .751)
> p.adjust(pvalues,method="bonferroni")
[1] 0.02 0.05 0.15 1.00 1.00 1.00 1.00 1.00 1.00 1.00
> p.adjust(pvalues,method="hochberg")
[1] 0.020 0.045 0.120 0.751 0.751 0.751 0.751 0.751 0.751 0.751
> p.adjust(pvalues,method="BH")
[1] 0.0200000 0.0250000 0.0500000 0.2825000 0.3783333 0.3783333 0.6485714
[8] 0.6900000 0.7366667 0.7510000
```

Other adjustments can be requested using "holm", "hommel", and "BY" (for the Benjamini, Hochber, and Yekutieli procedure).

Drug project example: Crestor (rosuvastatin)

- A commercial request was to compare rosuvastatin to other statins dose-to-dose.
- STELLAR was a 15-arm parallel group study comparing doses of rosuvastatin to doses of other statins: rosuva 10, 20, 40, 80 mg versus atorva 10, 20, 40, 80 mg versus prava 10, 20 40 mg versus simva 10, 20, 40, 80 mg. The primary variable was percent change from baseline in LDL-C.
- To address this objective, 25 pairwise comparisons of interest were specified.
- A Bonferroni correction was used to account for multiple comparisons.
- The sample size was estimated considering the Bonferroni correction. It was a large study, with about n=150 per arm.
- Choice of the conservative Bonferroni correction was influenced by the fact that a competitor received a warning letter from the FDA for dose-to-dose promotion from a study that was not designed to do dose-to-dose comparisons.
- There was no discussion with the FDA about correction for multiplicity in STELLAR. Results are considered robust, and they appear in the Crestor label.

References

1. Jones PH et al. Comparison of the efficacy and safety of rosuvastatin versus atorvastatin, simvastatin, and pravastatin across doses (STELLAR trial). Am J Cardiol 2003;92:152-160.

2. McKenney JM et al. Comparison of the efficacy of rosuvastatin versus atorvastatin, simvastatin, and pravastatin in achieving lipid goals: results from the STELLAR trial. Current Medical Research and Opinion 2003;19(8):689-698.