CHL 5225H
Advanced Statistical Methods for Clinical Trials: Multiplicity

Prof. Kevin E. Thorpe

Dept. of Public Health Sciences
University of Toronto

Objectives

1. Be able to distinguish among the various multiplicity issues present in clinical trials.
2. Understand why special care is called for when analysing trials where multiplicity is present.
3. Acquire a basic understanding of some of the methods commonly employed in dealing with multiplicity.

Types of Multiplicity

- comparisons among several (more than two) treatments
- multiple outcomes
- multiple analyses (eg. interim analyses) of a single outcome
- sub-group analyses

The Main Problem

Regardless of the context, when multiplicity exists, multiple analyses, usually in the form of hypothesis tests, are undertaken. Therein lies the potential to increase the Type I error.

Example

Suppose you are analysing two outcomes A and B in a clinical trial. Let E_A and E_B represent the events that a type one error occurs for events A and B respectively. Assume these events to be independent and that all analyses assume $\alpha = 0.05$. Then,

$$\begin{aligned} P(\text{Type I Error}) &= P(E_A) + P(E_B) - P(E_A E_B) \\ &= 0.05 + 0.05 - 0.05(0.05) \\ &= 0.0975 \end{aligned}$$

Comparing Several Treatments

- Suppose one has a clinical trial where more than two treatments are under study. It could be multiple doses of the same drug or completely different treatments. Often, it is desirable to examine all pairwise treatment differences.
- With three treatments you would have three comparisons, with four treatments you get six comparisons and with k treatment groups you get $\binom{k}{2} = k(k-1)/2$ comparisons.
- A large amount of work has been done in the context of continuous outcomes where one applies a multiple testing procedure following a statistically significant ANOVA.

Multiple Comparison Procedures

- We will consider the situation of a one-way ANOVA where the "treatment" factor has more than two groups.
- Many procedures have been proposed. We will consider four.
  - Bonferroni methods
  - Tukey's multiple range test
  - Scheffé's method
  - Dunnett's multiple comparison with a control
- Except for Bonferroni, the rest usually use a confidence interval approach.

One-way ANOVA

Recall the one-way classification model:

$$y_{ti} = \eta + \tau_t + \epsilon_{ti}$$

where $t = 1, 2, \ldots, k$, $i = 1, 2, \ldots, n_t$ and ϵ_{ti} are iid $N(0, \sigma^2)$. The usual ANOVA table is

Source	SS	DF	MS
Model	$S_T = \sum_t^k n_t(\bar{y}_t - \bar{y})^2$	$\nu_T = k - 1$	$s_T^2 = S_T/\nu_T$
Error	$S_R = \sum_t^k \sum_i^{n_t}(y_{ti} - \bar{y}_t)^2$	$\nu_R = N - k$	$s^2 = S_R/\nu_R$
Total	$\sum_t^k \sum_i^{n_t}(y_{ti} - \bar{y})^2$	$N - 1$	

Then to test the hypothesis

$$H_0 \ : \ \tau_1 = \tau_2 = \cdots = \tau_k$$

we compare the ratio $F = s_T^2/s^2$ to an F distribution with ν_T and ν_R degrees of freedom.

Bonferroni Methods

- The idea is simple. If you are making c comparisons (tests), in order to decalare a "statistically significant difference" you require $p \leq \alpha/c$.
- So, to compare a pair of means you would calculate a test statistic
$$t = \frac{\bar{y}_i - \bar{y}_j}{s\sqrt{1/n_i + 1/n_j}}$$
where s is the square root of the Error Mean Square from the ANOVA. The test statistic is compared to a t distribution with ν_R degrees of freedom.
- Although this tends to be too conservative when $c > 3$ it has widespread use for "ballparking" in many multiple testing situations.

Tukey's Method

In comparing k means, we wish to give confidence intervals for $\eta_i - \eta_j$ taking into account the fact that all possible comparisons may be made.

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_{k,\nu,\alpha}}{\sqrt{2}} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $q_{k,\nu,\alpha/2}$ is the appropriate upper significance level of the *studentized range* for k means, and ν is the number of degrees of freedom in the estimate s^2 of variance σ^2 (ie. the Error Mean Square from the ANOVA).

Scheffé's Method

This method computes confidence intervals for $\eta_i - \eta_j$ using the following

$$(\bar{y}_i - \bar{y}_j) \pm \sqrt{(k-1)F_\alpha(k-1,\nu)} s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where $F_\alpha(k-1,\nu)$ denotes the upper α point on the F distribution with $k-1$ and ν degrees of freedom.
This test is more conservative than others (except Bonferroni) but it is compatible with the overall ANOVA F-test. That is, it will not decalare a pairwise difference "significant" if the overall F-test is "non-significant."

Dunnett's Method

Suppose one of the treatments is a control (placebo) and the remaining treatments are active. One question of interest is whether any of the active treatments differ from the control. Let $\bar{y}_0$ be the observed average for the control group. We are now interested in confidence intervals on the $k-1$ differences $\bar{y}_i - \bar{y}_0$

$$(\bar{y}_i - \bar{y}_0) \pm t^*_{k,\nu,\frac{\alpha}{2}} s \sqrt{\frac{1}{n_i} + \frac{1}{n_0}}$$

where $t^*_{k,\nu,\alpha/2}$ is Dunnett's t.
It is good practice to allot more patients n_0 to the control group than to the active groups. As a guideline, $n_0/n_i \approx \sqrt{k}$.

Multiple Endpoints

- A clinical trial should have a single primary outcome.
- Usually, there are some additional secondary outcomes of interest to investigators. Some may be efficacy related while others may be safety related.
- There are various approaches that people have considered.

Composite Endpoints

- This is a common approach for combining multiple binary endpoints into a single binary endpoint.
- Let $y_1, \ldots, y_k$ be binary endpoints. The composite endpoint y is defined

$$y = \begin{cases} 1, & \text{if any } y_i = 1,\ i = 1, \ldots, k \\ 0, & \text{if all } y_i = 0,\ i = 1, \ldots, k \end{cases}$$

- However, it *must* make "sense" to combine endpoints.

Efficacy versus Safety

- Even if a single primary endpoint is defined for a trial, there may be many measures of treatment safety collected. They may be a mixture of continous (eg. lab values) and binary (eg. adverse events) and so cannot be combined. It may not even make sense to combine since a headache is of much less concern than a heart attack.
- The question of safety cannot be answered on statistical grounds alone.

Some Definitions

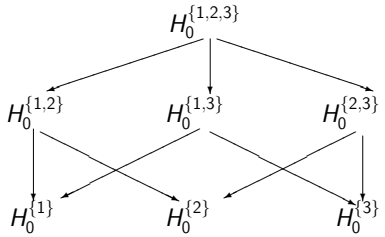
- The *Per Comparison Error Rate* (PCER) is the Type I error rate associated with each individual comparison.
- The *Experimentwise Error Rate* (EWER) is the probability that one or more of the single null hypthoeses that are true get rejected.
- The *False Discovery Rate* (FDR) is the proportion of errors among rejected hypotheses.

Bonferroni — Again

- The Bonferroni approach is often used with multiple endpoints. Suppose you are testing m endpoints. With the Bonferroni procedure you test each endpoint individually but only declare a statistically significant result for $p \leq \alpha/m$.
- This sets a PCER of α/m and ensures that EWER $\leq \alpha$. However, as stated earlier, this is too conservative when making many comparisons. In particular, it is not too helpful with genomic data.
- The advantage, of course, is its simplicity and the ability to consider different kinds of outcomes (eg. continuous, binary).

Closed Testing Procedures

- We need a system of hypotheses that is closed under intersection. The following figure illustrates a system with three endpoints.



- All hypotheses are tested at the global α level, but in order to test an hypothesis, all hypotheses higher in the hierarchy must have been tested and significant at level α. With this approach, the EWER will not exceed α. It is implied that there exist local α-level tests for each hypothesis in the hierarchy.

FDR Controlling Procedure

- Suppose we are testing a treatment effect in m endpoints by the hypotheses $H_1, \ldots, H_m$ yielding p-values $P_1, \ldots, P_m$. Let $P_{(1)} \leq \cdots \leq P_{(m)}$ be the ordered p-values and let $H_{(i)}$ be the null hypothesis corresponding to $P_{(i)}$.
- Let k be the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha^*$, then reject all $H_{(i)}$ $i = 1, \ldots, k$.
- Then the FDR will be controlled at α^*.
- An alternative definition has also been proposed for $\alpha^* = \alpha$, specifically, $P_{(i)} \leq \frac{i}{m+1-i}\alpha$.
- See Benjamini and Hochberg (1995) for details.

Multiple Analyses

- This occurs when a trial is analysed at various times as data accumulate, generally for the purpose of early termination due to efficacy or safety.
- There is a large body of literature on this subject under the heading of "Group Sequential Designs" and related terminology.

Reasons to Stop a Trial Early

- highly beneficial therapy — unethical to withhold a highly effective therapy
- harmful therapy — the opposite of a beneficial therapy
- futility — the trial is doomed to fail to reject the null hypothesis
- safety — excess toxicity, adverse events in one trial arm

The statistical/methodological approach will be dependent on the reason for the multiple analyses.

Stopping for Benefit/Harm

- Rationale:
  - If the experimental treatment under study in the trial is in fact beneficial/harmful, the trial team would like to know this as soon as possible so that the trial may be stopped and all appropriate patients may benefit or be protected.
- Issues:
  - Multiple looks at the primary outcome data are required.
  - Necessary to control Type I error.
  - Early termination reduces the precision of the treatment effect estimate and yields a biased estimate as well.
  - Stopping rules for benefit are primarily statistical in nature.

Stopping for Benefit/Harm — Continued

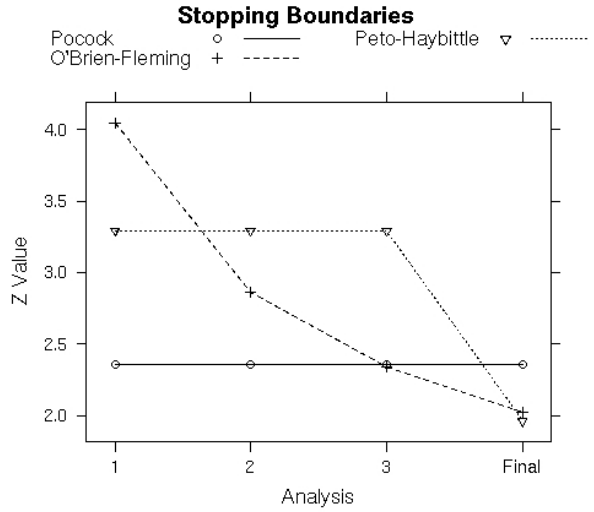
- Methodology:
  - The trial should be *designed* to accommodate multiple looks (interim analyses) using appropriate methods
    - group sequential designs
    - alpha spending functions
  - Interim analyses should be reviewed by a committee independent of the trial management team (in my opinion)

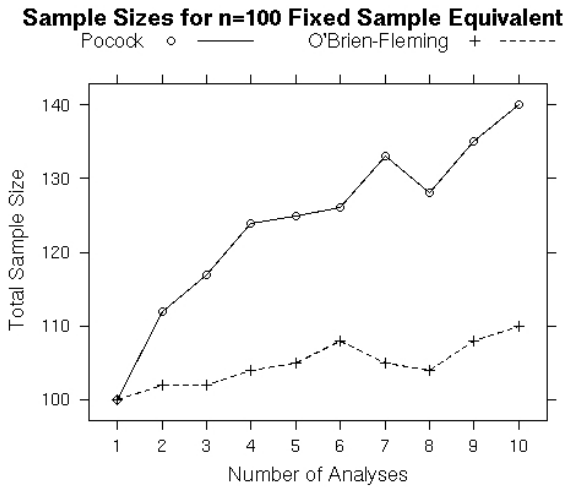
Group Sequential Designs

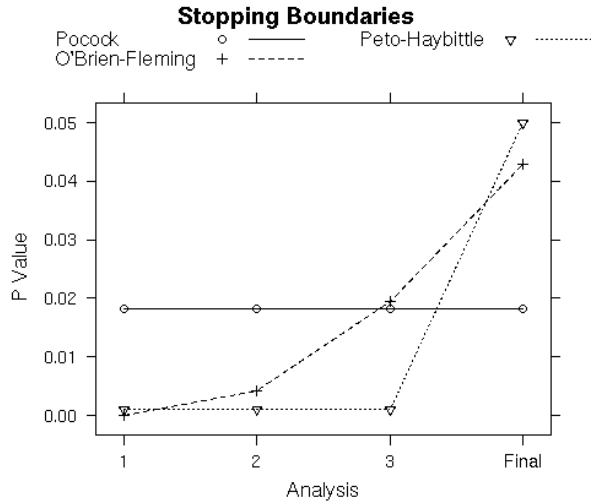
- A "small" number of interim analyses are planned to occur after pre-determined proportions of the anticipated data are available.
  - eg. 25%, 50%, 75% and 100% (normal end-of-study)
- Various stopping boundaries have been proposed.
  - Peto-Haybittle (interims at $p = 0.001$, final at $p = 0.05$)
  - Pocock's constant P ($p = 0.018$ with four analyses)
  - O'Brien-Fleming ($p_1 = 0.0001$, $p_2 = 0.004$, $p_3 = 0.019$, $p_4 = 0.043$)

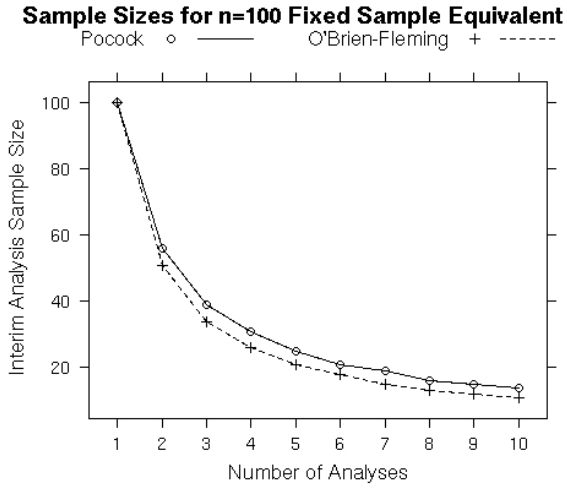
Group Sequential Designs (continued)

- If an interim analysis crosses the pre-specified boundary, a recommendation to stop would be expected, otherwise a recommendation to continue would be expected.
- Use of group sequential method generally has some effect on sample size calculations.


Stopping Boundaries
Pocock
O'Brien-Fleming
Peto-Haybittle
Z Value
4.0
3.5
3.0
2.5
2.0
1
2
3
Final
Analysis


Sample Sizes for n=100 Fixed Sample Equivalent
Pocock
O'Brien-Fleming
Total Sample Size
140
130
120
110
100
1
2
3
4
5
6
7
8
9
10
Number of Analyses


Stopping Boundaries
Pocock
O'Brien-Fleming
Peto-Haybittle
P Value
0.05
0.04
0.03
0.02
0.01
0.00
1
2
3
Final
Analysis


Sample Sizes for n=100 Fixed Sample Equivalent
Pocock
O'Brien-Fleming
Interim Analysis Sample Size
100
80
60
40
20
1
2
3
4
5
6
7
8
9
10
Number of Analyses

Pros and Cons

- Pros:
  - Relatively easy to design.
  - Simple to apply the stopping rules.
- Cons:
  - Since the completion of interim analyses depends on patient accrual and data collection, the actual timing of analyses is unpredictable and possibly inconvenient.
  - Changing the number of interim analyses once the trial begins is problematic.

Alpha Spending Functions

- The basic idea is that the significance level (p-value) that results in a recommendation for early termination is a function of the amount of *information* obtained to-date.
- This *spending function* describes the rate at which the total alpha is used up with accumulating data.
- Interim analyses can be planned at regular "time" intervals.
- "Extra" interim analyses can be accommodated.

Analysis Following Trial Termination

- In general, treatment effect estimates, confidence intervals and p-values require adjustment following completion of a trial with interim monitoring.
- Adjustments difficult to do by hand.
- If the trial stopped early, there could be substantial bias present in the estimate of treatment effect requiring correction.

Stopping for Futility

- Rationale:
  - If it becomes clear during the execution of the trial that there is virtually no hope to reject the null hypothesis at the end of the study, the trial team may wish to "cut their losses" and move on.
- Issues:
  - Multiple looks at primary outcome data.
  - Under some methods, Type I error is *not* affected however, Type II error may be affected.
  - A mixture of statistical and clinical judgement are often required.

Stopping for Futility — Continued

- Methodology:
  - Various statistical methods.
    - Conditional power.
    - Confidence intervals on treatment effect estimate.
    - Can be incorporated in some group sequential designs (eg. inner wedge designs).
  - Should be reviewed by an independent committee.

Conditional Power

Definition: The probability of the trial attaining "significance" at its final analysis point, given the current data.

Assumptions: Need to make assumptions about what you are likely to see in the future for the control group and the treatment effect.

Conditional Power Assumptions

- Control Group
  - Need to estimate the response we will see in the control group at the final analysis point.
  - Two sensible choices are the original expectation used in the sample size and a weighted average of the current response and the original assumption.
- Treatment Group/Effect
  - Need to estimate the response we will see in the treatment group at the final analysis point.
  - Best not to use the current treatment effect but rather impose the original planned treatment effect on the control group estimate.

Example

Two Group Binomial

Suppose a study was planned where the control group event rate was expected to be $\pi_c = 0.3$ and the treatment was expected to reduce this by $\Delta = 0.15$ resulting in a treatment group event rate of $\pi_t = 0.15$. Standard sample size formula yield $n = 134$ per group assuming $\alpha = 0.05$ (two-sided) and 80% power. A futility analysis was planned after half ($n = 67$ per group) the patients were "in" the trial.
Suppose the observed data were $x_c = 16$, $n_c = 67$, $x_t = 14$ and $n_t = 67$.
Use $\hat{\pi}_c = \pi_c = 0.3$ or

$$\hat{\pi}_c = \frac{n_c}{n}\left(\frac{x_c}{n_c}\right) + \left(1 - \frac{n_c}{n}\right)\pi_c \approx 0.27$$

and $\hat{\pi}_t = \hat{\pi}_c - \Delta$.

Example

Continued . . .

In the remainder of the trial, suppose the observed data are x'_c events in n'_c control patients and x'_t events in n'_t treatment patients.

The probability of observing x'_c and x'_t events in the remainder of the trial is obtained using the binomial distribution as follows.

$$P(x'_c, x'_t) = \left(\binom{n'_c}{x'_c} \hat{\pi}_c^{x'_c} (1-\hat{\pi}_c)^{n'_c - x'_c} \right) \left(\binom{n'_t}{x'_t} \hat{\pi}_t^{x'_t} (1-\hat{\pi}_t)^{n'_t - x'_t} \right)$$

The conditional power is then calculated as the sum of the probabilities for all combinations of future data $\{x'_c, x'_t\}$ which result in a statistically significant difference.

Example

Continued . . .

Thus,

$$\text{Conditional Power} = \sum_{|z|>1.96} P(x'_c, x'_t)$$

where,

$$z = \frac{(x_t + x'_t)/(n_t + n'_t) - (x_c + x'_c)/(n_c + n'_c)}{\sqrt{\hat{\pi}(1-\hat{\pi})(1/(n_t + n'_t) + 1/(n_c + n'_c))}}$$

and,

$$\hat{\pi} = \frac{x_t + x'_t + x_c + x'_c}{n_t + n'_t + (n_c + n'_c)}$$

If the conditional power is below some pre-specified value, you would recommend stopping for futility.

Example

Concluded

Returning to the example data, if we use $\hat{\pi}_c = 0.3$ we get a conditional power of about 39%.

If we use $\hat{\pi}_c \approx 0.27$ the conditional power is about 42%.

Some Comments

- In my experience, conditional power is most useful when asked to create and apply a stopping rule for futility in a trial already underway that was not designed with group sequential methods.
- If stopping for futility is likely to be an issue, design it in to the group sequential design in the first place.

Stopping for Safety

- Rationale:
  - If the experimental treatment results in an unacceptably high risk of "harm," the trial may need to be stopped (or modified) to prevent patients from being exposed to this risk.
- Issues:
  - Often there are a multiplicity of safety outcomes that are considered.
  - The primary outcome may also be a safety outcome (eg. mortality) in which case Type I error needs consideration.
  - Many safety outcomes are short-term where primary outcomes may be long-term.
  - Clinical judgement often plays a central role in interpreting safety data.
  - Need to carefully balance risk versus (potential) benefit to avoid "blowing the whistle" too soon.

Stopping for Safety

- Methodology:
  - Often less rigourously defined than benefit and futility rules.
  - Some aspects of safety may be covered within the group sequential design of the trial.
  - Confidence intervals of treatment effect are often useful as are composite safety outcomes to assess whether or not the safety profile is all one-sided.
  - Should be reviewed by an independent committee.

Sub-Group Analyses

- It is often the case with clinical trials that a large amount of data, besides the primary outcome, is collected. This includes demographic data (eg. age, gender) and clinical data that are thought to be related to the outcome(s) of interest.
- Investigators are then usually interested in not only which "extra" variables are associated with the outcome but whether or not different treatment effects are present for certain sub-groups of patients.
- This can result in a large number of tests and p-values.

Some Suggestions

- Usually, multiple regression models are employed here. Nevertheless, it can be tricky to disentangle real effects from spurious effects. Some of the ideas already encountered are often useful here:
  - Bonferroni methods.
  - False discovery rate methods.
- Another approach with sub-group analyses is to consider everything as hypothesis generating and use p-values as a guide to priority for additional study.

Interactions

- Interactions, especially with treatment, present a special problem since most trials are not "powered" to detect interactions. Before reporting that an interaction *may* be present, the following should be considered.
  - Does the interaction make biological sense?
  - Are the sub-group interaction effects clinically important?
  - The sample sizes of the sub-groups should not be extremely imbalanced to provide some robustness.
  - The p-value should be very small to be more confident that the effect *might* be for real.

Additional References

1. *Statistics for Experimenters*; Box, Hunter and Hunter, Wiley, 1978, Appendix 6C.
2. Scheffé, *A Method for Judging all Contrasts in the Analysis of Variance*; Biometrika, Vol. 40, No. 1/2 (Jun., 1953), 87–104.