Multiple Tests, Interim Monitoring, Subset Analyses

- These activities all have a common element of conducting and then interpreting multiple statistical tests
- Begin discussion with a general overview.
- Then: Interim monitoring
- Then: Subset analysis

Multiple Tests and Type I Error

- Often situations when multiple tests are conducted that all involve the same statistical hypothesis
  - Multiple interim analyses
  - Comparing treatment A and B in various subsets of the entire data (eg, among males, among patients with prior treatment, etc)
  - Using different metrics for the outcome variable (eg, change from baseline to week 2, change from baseline to week 3, ..)

Multiple Tests and Type I Error (cont)

- Well known that the overall chances of a Type I (false positive) result are inflated when multiple tests are carried out without 'adjustment'
- Example: 10 independent statistical tests, each at the 0.05 level of significance. Suppose no treatment difference at the population level:
  - P(any particular test will be 'signifiicant')=5% (usual situation)
  - P(at least 1 of the 10 will be 'significant')>40%
  - Prob(at least 1 False Pos Result) >40%
  - Suppose we did 10 tests, and 1 was 'significant'. Then reporting this test without context of setting (just 1 of 10) is misleading

Multiple Tests and Type I Error (cont)

- Suppose we want to adjust statistical tests so that overall false positive rate is 5%.
- Several ways to do this, but can be complicated when tests are correlated:
  - Bonferroni correction-can be very conservative.
  - Formally account for the correlation among tests-can be complex except in specific situations (interim monitoring)
  - Limit number of tests conducted
    - Either literally, or by prespecifying 1 or 2 to be 'primary endpoints' and others to be 'secondary' or 'tertiary' endpoints. Then, place main interpretation on primary endpoints.
  - In all cases, be clear about what was examined

Next….Interim Monitoring of Trial Results

- Common to review interim results of a trial periodically (often, every year)

- Review often done by an independent committee-Data and Safety Monitoring Committees---”DSMC” (or DMC, DSMB)

- DSMC recommends if the trial should be modified or terminated prior to its planned completion

Data and Safety Monitoring Committees

- Composition
  - biostatisticians, clinicians, ethicists, other scientists (sometimes), patient advocates (sometimes)
- Tasks
  - review study conduct
  - review safety and toxicity data
  - review interim efficacy data
- Actions
  - recommendation on continuation of study
  - recommendation on modifications

Rationale for Interim Monitoring

Once a trial has begun, it should be continued only if

- It remains ethical to randomly assign the study treatments
- The study continues to have the potential to achieve its scientific goals

Reasons for Stopping a Trial During an Interim Analysis

- Treatments are convincingly different *
- Treatments are convincingly not different *
- Unacceptable side effects or toxicity
- Accrual is so slow that trial is no longer feasible
- External information makes the trial unnecessary or unethical
- Poor execution compromises the ability of the study to meet its objectives
- Catastrophic fraud or misconduct

Group Sequential Methods

- Most common approach used in RCTs today
- Before trial begins: plan to examine the interim data K times; e.g.
  - every year for K=3 years
  - after 100, 150, and 200 patients have been enrolled and followed for 6 months (K=3)
- At each interim analysis, compare treatments and decide whether there is sufficient evidence to stop the trial (eg, because new treatment is better than standard)

- Null Hypothesis: H_0: Trmt A=Trmt B
- Let Z_k denote the test statistic we use to compare treatments during the k-th analysis, and we have ‘stopping boundaries’ $B_1, B_2, \ldots, B_K$ that govern stopping of the trial:
  - If |Z1|>B1, stop the trial in favor of A (if $Z_1 < -B_1$) or B (if $Z_1 > B_1$); otherwise continue to 2nd analysis
  - If $|Z_2| > B_2$, stop the trial in favor of A (if $Z_2 < -B_2$) or B (if $Z_2 > B_2$); otherwise continue the study to the 3rd analysis
  - Continue until end (K-th analysis); If the trial is completed without exceeding the boundaries, do not reject H_0

Study Design with Interim Monitoring

- We want an adequate sample size to achieve a desired power for a particular alternative and a desired Type 1 error (usually 5%).
- Type I error in this context=

  P(declare a treatment difference | H0).
- In our group sequential design, we will declare a difference between groups if
  - |Z1|>B1 (stop at 1st interim),
  - |Z1<B1| but |Z2|>B2 (stop at 2nd interim),
  - …
  - |Z1|<|B1|, |Z1|<B2, …, $|Z_{k-1}|<B_{K-1}$, |ZK|>BK (last analysis)

Determining the Boundaries $B_1,\ldots, B_K$

- Want to choose B1,…,BK so that under H0, .05=P[reject H0 at one of the K analyses]

- If we denote P[reject H) at kth analysis] by

π_k , then we want (mutually exclusive events)

$$\pi_1 + \pi_2 + \ldots + \pi_K = .05$$

- Many ways to choose the π_k. These determine at what rate we ‘spend’ the .05 type I error.

Stopping Boundaries

- Pocock was the first to take this approach and chose equal π_k .For example, if K=3, each $\pi_k = .05/3=.0167$
- Most popular method today is probably one proposed by O'Brien & Fleming. They 'spend' very little of the Type I error at beginning interim analyses, making it harder to stop early. However, in return, little adjustment needed at end.

Critical Values -Nominal P values: K=4 analyses, Type I error = 0.05

	Pocock		O’Brien-Fleming	
Analysis	CV	P	CV	P
1	2.36	.016	4.08	.000005
2	2.36	.016	3.22	.0013
3	2.36	.016	2.28	.0228
4	2.36	.016	2.04	.0417

(CV=critical value for Z statistic)

Choosing Boundaries

- Pocock boundaries: greater chance of stopping the trial early
- O’Brien-Fleming: more difficult to stop early, and thus p-values at end close to nominal (0.05) level

- Note: different rules on how to report p-values when a trial is stopped early; not discussed here.

Modified Group Sequential Procedures

- Stopping rules can be modified to allow
  - interim analyses at uneven intervals
  - Stopping the trial for “futility” ”—that is, if the chances of finding a difference in the future is sufficiently low

Example: Stopping for Futility

- Reference: Hall et al, NEJM, 1998, 338: 1345-51
- Treatment of Progressive Multifocal Leukoencephalopathy (PML) in patients with HIV receiving ART:
  - IV Cytarabine
  - Intrathecal Cytrabine
  - Control.
- Main endpoint: survival
- N=57

Treatment of PML

- Interim analyses planned; stochastic curtailment used to compute the conditional power of the study, given interim results
- At 2nd interim analysis:
  - 14 deaths in each group (p=0.85)
  - Conditional probability of finding a significant difference among the 3 groups, IF THE TRIAL WERE COMPLETED, was less than 1%.
- Trial was terminated. Lack of evidence to support use of Cytarabine

Treatment of PML

- Stopping a trial early for futility does not impact Type I error.
- Stopping for futility does not mean that new treatment has been shown to be no better, only that continuation of trial is unlikely to demonstrate a difference.
- In PML example, no evidence of a survival difference when about 60% of the planned information was available. Little to argue for its use in practice.

Next… Multiple Statistical Tests
Selective Reporting and Subset Analyses

- Inflated Type I error due to multiple testing can also arise when conducting multiple statistical tests of the same endpoint, or when conducting tests on multiple subgroups of the patient population. We illustrate this with 2 exmples:
- Example 1: HIV/AIDS Study: Not reporting all of the facts
- Example 2: Lung Volume Reduction
  Surgery Trial: Subset analyses

Subset Analyses

- Common Situation: Conduct trial, no significant difference between treatment groups overall. Then begin to examine treatment differences in patient subgroups, eg:
  - Among male patients
  - Among older patients
  - Among older male patients
  - Among patients with prior treatment
  - ….etc
- Nothing wrong in doing this !!
- Problems arise when p-values are not adjusted to correct for the inflated Type I error due to multiple testing

Correcting for Multiple Tests

- complicated because tests are usually correlated:
  - Bonferroni correction-can be very conservative.
  - Limit number of tests conducted
    - Either literally, or by prespecifying 1 or 2 to be ‘primary endpoints’ and others to be ‘secondary’ or ‘tertiary’ endpoints. Then, place main interpretation on primary endpoints.
  - In all cases, be clear about what was examined. “We examined 25 subsets of the population, and found 3 where the treatment appears to show some suggestion of benefit …”

Example 1: Not Reporting All Tests

- References:
  - Churnboonchard et al. Clin. Diag. Lab. Imm. (2000) 7: 728-733.
  - Glidden et al. Clin. & Diag. Lab. Imm. (2001) 8:468-69. (Letter)
- HIV AIDS trial. Comparing new therapy versus placebo. Patients evaluated multiple times during the study
  - Main endpoint: Changes in CD4
  - Prespecified primary analysis: comparison of slope of log(CD4) between groups using nonparametric test
  - Multiple secondary analyses using other metrics for change in CD4

Results of Prespecified Primary and Secondary Statistical Analyses of CD4 Count

	Method for	log-transformed		
	Calculating CD4	**CD4 counts?**	**Metric**	**P-value***
1.	original	yes	slope	0.34 **
2.	recalculated	yes	slope	0.36
3.	original	yes	change by wk 40	0.34
4	recalculated	yes	change by wk 40	0.20
5.	original	no	change by wk 40	0.13
6.	recalculated	no	change by wk 40	0.07
7.	original	yes	AUC	0.11
8.	recalculated	yes	AUC	0.07
9.	original	no	AUC	0.044
10.	recalculated	no	AUC	0.024

* not adjusted for multiple comparisons

** pre-specified primary analysis

Example 1 (continued)

- Problems:
  - Authors reported only the 10th (most significant) analysis
  - Did not report that this was not prespecified analysis
  - Did not report that there were 10 analyses done
- Misleading to reader—exaggerates the evidence in favor of new treatment

Example 2:Lung Volume Reduction Surgery vs Med Treatment for Severe Emphysema

- Reference: NEJM, 2003, 348:2059-73
- N=1218 patients with severe emphysema. After pulmonary rehabilitation, patients randomized to:
  - A: lung volume reduction surgery (608), vs
  - B: medical treatment (610)
- Primary endpoint: death
- Secondary endpoints: QoL, exercise capacity

Lung Volume Red. Surgery vs Medical Treatment: Results for all Patients

	LVRS	MedT	p-value
• Deaths:	157	160	0.90
• Improved Ex. Capacity*	54	10	<0.001
• Qual. Life*	121	34	<0.001

*evaluated in 371 (LVRS) and 378 (MT) patients

Lung Volume Red. Surgery vs Medical Treatment: Subgroup Analyses

- Surgery vs Medical modalities compared in multiple patient subgroups, and two suggested an interaction with treatment group w.r.t. mortality
  - Primary lobe of emphysema: upper vs non
  - Exercise capacity: low vs. high
- These 2 factors then used to divide patients into 4 subgroups:
  - Upper lobe & low exer. capacity
  - Upper lobe & high exer. capacity
  - Non upper lobe & low ex. capacity
  - Non upper lobe & high ex. capacity

Lung Volume Red. Surgery vs Medical Treatment: Subgroup Analyses

- Mortality differences between Surgery and Medical Treatment groups suggested within these 4 subgroups
  - Surgery better: Upper lobe, low exercise capacity: p=.005
  - Medical Trmt better: Non-upper lobe, high exercise capacity: p=0.02
  - No significant difference: other 2 subgroups

Lung Volume Reduction Surgery vs Medical Treatment: Overall Interpretation

- Overall:
  - No significant mortality difference
  - Improved exercise capacity and QoL in surgery group
- Multiple subgroups examined, suggested mortality difference when 2 are combined
  - Surgery better: upper lobe, low exercise capac.
  - Medical better: non-upper lob, high exercise capac
- Subgroup results are plausible, but is the evidence definitive?
- Should medical practice change?

Lung Volume Red. Surgery vs Medical Treatment

- This is an example where authors tried to assess subgroups in a responsible way, and did provide some caution in their conclusions due to subgroup analyses

- While the subgroup results are not definitive, they provide evidence that can be helpful in guiding physicians & patients with severe emphysema in their treatment options

Summary

- Multiple Tests inflate the chances of a false positive finding.
- This needs to be recognized and, whenever feasible, accounted for by adjusting p-values. Group sequential methods represent one setting where this can be done.
- Subgroup analyses need to be done with caution. Evidence of a treatment difference ONLY in a subgroup should always be viewed with caution. Supportive evidence with secondary endpoints can be useful.
- When reporting results, do not just present the ‘good news’, but be clear about the number of types of analyses that were conducted, so that the reader can judge for himsel/herself the strength of the evidence