MULTIPLE COMPARISONS IN CLINICAL TRIALS

Susan S. Ellenberg, Ph.D.
Professor of Biostatistics
Perelman School of Medicine
University of Pennsylvania

May 2, 2018

POLLING QUESTION 1: WHO'S IN THE AUDIENCE TODAY?

1. Clinical researcher
2. Statistician
3. Clinical research staff
4. Student
5. Other

POLLING QUESTION 2: HAVE YOU BEEN INVOLVED IN A RANDOMIZED CLINICAL TRIAL?

1. Never
2. Yes, but only to enter patients
3. Yes, I have designed and conducted one or more clinical trials

THE PROBLEM OF MULTIPLICITY

- Multiplicity refers to the multiple judgments and inferences we make from data
  - hypothesis tests
  - confidence intervals
  - graphical analysis
- Multiplicity leads to concern about inflation of Type I error, or false positives

MULTIPLICITY IN CLINICAL TRIALS

- There are many types of multiplicity to deal with
  - Multiple endpoints
  - Multiple subsets
  - Multiple analytical approaches
  - Repeated testing over time

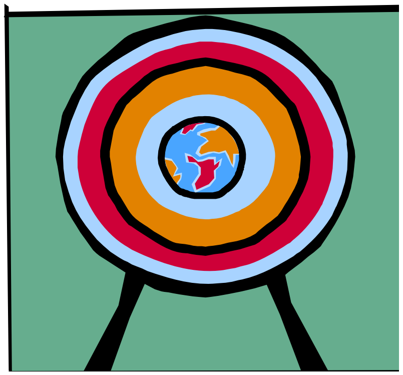
POLLING QUESTION 3: HOW MUCH THOUGHT HAVE YOU GIVEN TO MULTIPLICITY ISSUES?

1. I've never heard of this issue
2. I'm somewhat familiar with it but never had to deal with it
3. I'm reasonably familiar with it and have had to address it in my work
4. I'm an expert

MOST LIKELY TO MISLEAD: DATA-DRIVEN TESTING

- Perform experiment
- Review data
- Identify comparisons that look "interesting"
- Perform significance tests for these results

CHANCE OF A BULLS-EYE?



CHANCE OF A BULLS-EYE?

CHANCE AND COINCIDENCE

- A “false positive” is essentially a chance finding—a coincidence
- We often marvel at coincidences without recognizing how many opportunities there are to observe such an event
  - The coincidence you observe is not the only one you MIGHT have observed

EXAMPLE

- The chance of drawing the ace of clubs by randomly selecting a card from a complete deck is 1/52
- The chance of drawing the ace of clubs at least once by randomly selecting a card from a complete deck 100 times is….?

EXAMPLE

- The chance of drawing the ace of clubs by randomly selecting a card from a complete deck is 1/52
- The chance of drawing the ace of clubs at least once by randomly selecting a card from a complete deck 100 times is….?
- And suppose we pick a card at random and it happens to be the ace of clubs—what probability statement can we make?

YOUNG'S FALSE POSITIVE RULES

- With enough testing, false positives will occur
- Internal evidence will not contradict a false positive result (i.e.,--don't imagine you'll be able to figure out which are the false positives)
- Good investigators will come up with a possible explanation
- It only happens to the other person

Westfall and Young, Resampling-Based Multiple Testing, John Wiley & Sons, 1993

COMMON PHRASES RELATED TO THE MULTIPLICITY PROBLEM

- Testing to a foregone conclusion

COMMON PHRASES RELATED TO THE MULTIPLICITY PROBLEM

- Testing to a foregone conclusion
- Data dredging

COMMON PHRASES RELATED TO THE MULTIPLICITY PROBLEM

- Testing to a foregone conclusion
- Data dredging
- Torturing the data until they confess

COMMON PHRASES RELATED TO THE MULTIPLICITY PROBLEM

- Testing to a foregone conclusion
- Data dredging
- Torturing the data until they confess
- P-hacking

A LONG-STANDING ISSUE THAT IS STILL OFTEN IGNORED

- Researchers
- Journal editors
- Reporters
- Consumers

WHY IS IT IGNORED?

- Clinical trials are expensive—need to learn as much as we can from each trial
- Adjusting for multiplicity means we do each test at reduced (often substantially reduced) significance levels—lose power
- Adjustment procedures can be very conservative when variables are correlated
- Reporting adjusted p-values may be confusing to readers
- No real consensus about what or how to adjust

ONE AREA WITH BROAD CONSENSUS: INTERIM ANALYSES

- Most researchers now recognize that regular interim analysis of emerging results as the trial progresses, with a strategy of stopping as soon as “p<0.05” is observed, will increase risk of false positive error
- Statistical boundaries to guide early termination considerations are widely used

FALSE POSITIVE RATES WHEN TESTING MULTIPLE TIMES AT THE 0.05 LEVEL

significance level (%)	No. of repeated tests						
	1	**2**	**3**	**4**	**5**	**10**	**200**
1	1	1.8	2.4	2.9	3.3	4.7	12.6
5	**5**	**8.3**	**10.7**	**12.6**	**14.2**	**19.3**	**42.4**

McPherson K, *New England Journal of Medicine*; 290:501-2, 1974

MANY OPPORTUNITIES FOR MULTIPLICITY

EXAMPLE: ONCOLOGY TRIAL

- Experiment : regimens A, B and C are compared to standard tx
  - Intent: cure/control cancer
  - Eligibility: non-metastatic disease

MANY OPPORTUNITIES FOR MULTIPLE TESTING

- Multiple treatment arms: A, B, C
- Subsets: gender, age, marker levels…
- Site groupings: country, type of clinic…
- Covariates accounted for in analysis
- Repeated testing over time
- Multiple endpoints
  - different outcome: mortality, progression, response
  - different ways of addressing the same outcome: different statistical tests

COMMON SCENARIOS THAT RAISE CONCERNS ABOUT THE ROLE OF COINCIDENCE

- No overall treatment effect, but effect seen in a subset: e.g.,
  - women
  - those over age of 50
  - those with less severe disease
  - (those who complied with protocol)
- No overall treatment effect at specified time point, but effect seen at earlier or later time point

FOUR BASIC APPROACHES TO MULTIPLE COMPARISONS PROBLEMS

1. Ignore the problem; report all interesting results
2. Perform all desired tests at the nominal level (e.g., p=0.05) and warn reader that no accounting has been taken for multiple testing—let readers "mentally adjust" as they will
3. Limit yourself to only one test
4. Adjust the p-values/confidence interval widths in some statistically justifiable way

IGNORE THE PROBLEM

- Probably the most common approach
- Less common in the higher-powered journals, or journals where statistical review is standard practice and they won't let you get away with it
- Even when not completely ignored, often not fully addressed

DO ONLY ONE TEST

- Single (pre-specified) primary hypothesis
- Single (pre-specified) analysis
- No consideration of data in subsets

- Not really practicable

NO ACCOUNTING FOR MULTIPLE TESTING, BUT REPORT APPROACH

- Message is that readers should “mentally adjust”
- Justification: allows readers to apply their own preferred multiple testing approach
- Appealing because you show that you recognize the problem, but you don’t have to decide how to deal with it
- May expect too much from statistically unsophisticated audience—but it’s easy

USE SOME TYPE OF ADJUSTMENT PROCEDURE

- Divide desired significance level α by the number of comparisons (Bonferroni)
- Bonferroni-type stepwise procedures
- Control false discovery rate

- Multivariate testing for heterogeneity, followed by pairwise tests

----

- Resampling-based adjustments
- Bayesian approaches

BONFERRONI ADJUSTMENT

- Early and still common approach
- Provides upper bound for false positive error
- Conservative when comparisons are correlated (non-independent)
- Will severely reduce power when many comparisons are made

BONFERRONI ADJUSTMENT

- Divide significance level by number of comparisons you want to make
- If you have 2 main endpoints and want to declare a positive result if you show a statistically significant difference on either, need to test each at 0.025
  - If 5 endpoints, test each at 0.01
  - Can continue to divide p-value for other testing of interest

BONFERRONI ADJUSTMENT

- Statistically appropriate when comparisons are independent
  - Comparing results in separate populations
  - Comparing results on unrelated outcomes
- Conservative when comparisons are not independent
  - Measurements of same outcome at different time points
  - Use of different diagnostic criteria for same outcome
  - Analyses adjusted for different sets of covariates

INDEPENDENT vs NON-INDEPENDENT

- Independent tests
  - In a randomized trial conducted at 5 sites, a test for treatment effect at each site
- Non-independent tests
  - In a randomized trial of a treatment for pain relief, a test for differences in need for rescue medication, and a test for differences in pain scores

MODIFICATIONS OF BONFERRONI THAT ARE SOMEWHAT LESS CONSERVATIVE

STEPWISE BONFERRONI

- Holm (1979)
  - Use Bonferroni bound for test that produces smallest p-value
  - Can use successively less restrictive bounds for successive tests
- Simes (1986)
- Hochberg (1988)

STEPWISE BONFERRONI

- Suppose you had 6 primary/secondary hypotheses, and the p-values were as follows: 0.07, 0.009, 0.28, 0.017, 0.032, 0.0008.
  - Step 1: order p-values from largest to smallest: 0.28, 0.07, 0.032, 0.017, 0.009, 0.0008
  - Step 2: divide α by 6: 0.05/6 = .00833
  - Step 3: see if your smallest p-value is less than 0.00833 (It is! Can reject this hypothesis and continue)
  - Step 4: divide α by 5: 0.05/5 = 0.01
  - Step 5: see if your next-smallest p-value is less than 0.01 (It is! Can reject this hypothesis also)
  - Step 6: divide α by 4: 0.05/4 = 0.0125
  - Step 7: see if your next smallest p-value is less than 0.0125 (sorry, no, so stop and fail to reject remaining hypotheses)

A DIFFERENT APPROACH

- Bonferroni and its variations control the Familywise Error Rate (FWER)
  - Focus is on limiting the probability of making <u>any</u> type 1 error
- Benjamini and Hochberg developed an approach controlling the False Discovery Rate (FDR)
  - Focus is on limiting the proportion of type 1 errors among all hypotheses tested

CONTROLLING FALSE DISCOVERY RATE

- Benjamini and Hochberg (1995, *JRSS B*)
- New approach: controlling the expected proportion of false positives
- Procedure:
  - Define m hypotheses
  - Arrange m observed p-values in ascending order
  - Let k be largest i for which $P_{(i)}<=0.05i/m$
  - Can reject at the 0.05 level all null hypotheses with p-values less than or equal to $P_{(k)}$
- Maintains power at a higher level compared to other approaches, especially when many tests are to be performed

BENJAMINI/HOCHBERG APPROACH

- Take previous example: 6 p-values, 0.0008, 0.009, 0.017, 0.032, 0.07, 0.28
- Let i designate the order of the p-values
- Let k be largest i for which $P_{(i)}$<=0.05i/m, where m = 6
  - P_1 = 0.0008 < 0.05/6 (=.0083)
  - P_2 = 0.009 < 0.05x2/6 (=0.0167)
  - P_3 = 0.017 < 0.05x3/6 (=0.025)
  - P_4 = 0.032 > 0.05x4/6 (=0.03)

  - P_5 = 0.07
  - P_6 = 0.28
- Conclusion: can reject the first 3 null hypotheses at the 0.05 level (one more than with the Holm method)

CONCLUSIONS DIFFER

- Holm: reject 2 hypotheses and are assured that you have no more than a 5% chance of a type 1 error
- Benjamini/Hochberg: reject 3 hypotheses and are assured that no more than 5% of your null hypotheses are erroneously rejected
- Note: standard Bonferroni correction would have permitted rejection of only one hypothesis

PROCEDURES ARE STILL CONSERVATIVE

- They are still based on the assumption of independent comparisons
- If your comparisons are correlated, you'll still be overly conservative using any of these methods
- (But if your comparisons are highly correlated you'll probably meet the cutoff criteria)

MULTIPLE CHOICE

Which of the following is incorrect?

1. Multiplicity issues are often not addressed in reports of clinical trials
2. There is fairly broad consensus on the best way to handle the multiplicity issue
3. Adjusting statistical tests for multiplicity affects the power of the trial to detect treatment effects
4. Bonferroni corrections are often overly conservative

MULTIPLE CHOICE

Which of the following is incorrect?

1. Multiplicity issues are often not addressed in reports of clinical trials
2. **There is fairly broad consensus on the best way to handle the multiplicity issue**
3. Adjusting statistical tests for multiplicity affects the power of the trial to detect treatment effects
4. Bonferroni corrections are often overly conservative

SELECTING CUTPOINTS: ANOTHER PITFALL

- Many subsets are based on categorizing continuous endpoints
- How do we decide where to cut?
  - Clearly plausible threshold
  - Median of observed measures
  - Point that best divides group prognostically
  - Point that best divides group in regard to response to treatment

EXAMPLES

- Age
  - Some clearly plausible cutpoints: ages that define infancy, toddlerhood, adolescence, adulthood, female fertility
  - In adult populations, divisions often arbitrary (e.g., why 50 and over?)
- Smoking
  - How many cigarettes/packs per day?

SELECTING CUTPOINTS

- In some cases, standard or obvious categories
  - Laboratory values (normal, abnormal)
  - Apgar scores (0-3, 4-6, 7-10)
- In many cases, investigators may look for cutpoint that maximizes difference in outcomes between categories
- Often, authors will not explain how they selected the cutpoints

CUTPOINTS FOR PROGNOSIS

- Altman et al* looked at results of selecting a cutpoint corresponding to the most highly significant association with outcome
- Considered data from numerous studies on prognostic value of SPF (% of tumor cells in DNA-synthesizing phase obtained by cell-cycle analysis) in breast cancer
- Wide range of values used to define “high” and “low”

*Altman D, et al. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *JNCI*, 1994

ALTMAN ET AL: RESULTS

- Process of seeking the cutpoint that minimized p-value for significance of SPF in log rank analysis led to finding of cutpoints that appeared significantly prognostic
- Showed by simulation that when covariate has no prognostic value, can find a cutpoint making covariate appear "significantly" prognostic 40% or more of the time

KEY POINTS

- Be aware of problem
- Confirmatory analyses can only be performed when outcome and analytic approach has been pre-specified
- No single universally accepted way of dealing with multiple comparisons
- Optimal approach may differ in different situations
- Still area of active research

RESULTS IN SUBSETS

MOTIVATION FOR SUBSET ANALYSIS

- Not at all implausible that treatment might have varying effect in different subgroups
  - General prognosis
  - Co-existing conditions
  - Age, gender
  - Prior therapy
  - Genetic characteristics
- Physicians want to optimize approach for individual patients

Subset analyses are important in developing information about optimal treatment strategies

BUT

Subset analyses may be unreliable since multiple analyses frequently produce spuriously positive (or negative) results

FAMOUS EXAMPLE

- ISIS-2: major trial of antithrombotic therapy for MI
- Accepted for publication in *Lancet*
- Editors of *Lancet* wanted authors to include results in subsets
- Authors were skeptical of these results, but *Lancet* insisted
- Authors acquiesced, but added their own subsets in addition to those requested by *Lancet*

ISIS-2 RESULTS

- Overall results highly significant—multiple zeros to right of decimal pt
- Authors considered results according to zodiac sign under which subject was born
- Subset of subjects born under all signs except Libra and Gemini showed highly positive effects: 28% mortality reduction, p<0.0001
- Effects for those born under Libra or Gemini went in wrong direction: 9% mortality increase (not significant)

Subgroup analysis: a machine for generating false negatives

-- Richard Peto

ISIS-2	ASA	Placebo	RR	*P*
Gemini or Libra	11.1%	10.2%	1.09	NS
Others	9.0%	12.1%	0.72	<0.00001

Lancet. 1988 (slide borrowed from Rob Califf, Duke)

MULTIPLE CHOICE

Suppose a clinical trial compares two treatments and the underlying truth is that there is no difference in effect on outcome. Suppose there are 10 clinical sites, each with about the same number of participants. What is the probability that you will find a statistically significant difference in treatment effect in at least one site?

1. 5%
2. 20%
3. 40%
4. 60%

MULTIPLE CHOICE

Suppose a clinical trial compares two treatments and the underlying truth is that there is no difference in effect on outcome. Suppose there are 10 clinical sites, each with about the same number of participants. What is the probability that you will find a statistically significant difference in treatment effect in at least one site?

1. 5%
2. 20%
3. **40%**
4. 60%

PROBABILITY OF SPURIOUS RESULT (independent subsets)

K	Probability
2	0.10
5	0.23
10	0.40
20	0.64

AVOIDING SPURIOUS SUBSET FINDINGS

- Can test whether we have a statistically significant treatment by covariate interaction

TREATMENT BY COVARIATE INTERACTIONS

- We have an interaction between treatment and a covariate when the treatment effect depends on the value of the covariate
- There are statistical tests to assess the likelihood that an observed effect difference by covariate categories is real
- Examples of covariates that are known to affect treatment responsiveness
    - Estrogen receptor status, breast cancer
    - KRAS mutation, colorectal cancer
    - Age, influenza vaccination

TESTING FOR TREATMENT-COVARIATE INTERACTION

- Can test whether results in subgroups differ to a significant extent
- If interaction is significant, maybe looking at results in subsets is more defensible?
- Problem: power for such tests is low when trial powered on main effect
- Ad hoc approach that is commonly used: test at 0.20 level
- If many covariates to consider will have to do many interaction tests

COMMENT ON OBSERVATIONAL STUDIES

- Many observational studies are designed to address multiple hypotheses, not necessarily foreseen at time of initiation
- Selection bias and confounding are always present and of potentially great magnitude
- Relative risks of less than 2 (or maybe even 3 or 4) found in observational studies should be viewed skeptically, no matter how many zeroes follow the decimal point in the p-value

NO ARGUMENT AGAINST EXPLORING DATA!

- Clinical research is expensive
- Appropriate for researchers to explore data thoroughly, looking for clues to improved use of treatments
- Inappropriate to view such exploratory analyses as definitive; such clues require confirmation
- Sir Richard Peto: One should always do subset analyses, but never believe them

WOMEN'S HEALTH INITIATIVE

- Huge clinical trial of various interventions in postmenopausal women (ages 50-79)
- One substudy (of many): efficacy of calcium with vitamin D supplementation for preventing fractures
- Over 36,000 women randomized
- Primary hypothesis: Calcium+D will reduce rate of hip fracture
- Secondary hypothesis: Calcium+D will reduce rate of all fractures

RESULTS OF SUBSTUDY

- Primary outcome: suggestive but not significant decrease in hip fracture
- Secondary outcome: 4% (nonsignificant) decrease in total fractures
- Secondary outcome: small (6%) but significant increase in hip bone density
- Moderate and significant increase (17%) in kidney stones

WHI CATEGORIZED VARIABLES

- Age (50-59, 60-69, 70-79)
- Weight (over or under 58 kg)
- BMI (<25, 25-29, 30 and over)
- Smoking (never or past, current)
- Solar irradiance (5 groups based on Langleys)
- Physical activity (0-3, 3-11.75, >11.75 MET)
- Total calcium intake (<800, 800-1200, >1200)
- Total vitamin D intake (<200, 200-400, 400-600, >600)

SUBSETS ASSESSED FOR EFFICACY

- Age (3)
- Race/ethnic grp (6)
- Weight (2)
- BMI (3)
- Smoking status (2)
- Langleys (5)
- Falls in past yr (4)

- Physical activity (3)
- Prior fracture (2)
- Total Ca/day (3)
- Total Vit D/day (4)
- History of HT use (3)
- Grp in HT trial (2)

ADDITIONAL ANALYSES

- Bone mineral density
- Fracture site
  - Hip
  - Clinical vertebral
  - Lower arm or wrist
- Adherence
  - Followup censored 6 months after determination of nonadherence

RESULTS OF SUBGROUP ANALYSES

- Of the 13 variables considered, significant interaction of calcium/D treatment with respect to hip fractures reported for 2, nearly significant for another
- Subgroup analyses shown for only hip fractures, not 3 other fracture outcomes
- Authors note that up to 3 statistically significant interactions, considering each of 4 fracture outcomes, would be expected by chance

WHI ILLUSTRATES DIFFICULTIES WITH MULTIPLICITY

- Investigators were clearly aware of the issue and tried to address it
- Because of the importance of the study and the resources poured into it, investigators clearly wanted to explore the data thoroughly
- They tried to strike a balance between providing information and over-interpreting

WAYS TO LIMIT MULTIPLICITY PROBLEMS

- Define a primary hypothesis, with a specific analytic procedure
- Define a small number of secondary hypotheses, including any subset analyses of particular importance
- Consider using a statistical procedure that adjusts for multiple tests when outcomes are not highly correlated
- Perform (and report) tests of treatment by covariate interaction when subsets defined by the covariate are reported
- Avoid interpreting any analysis other than the primary analysis as “definitive”

CONCLUDING COMMENTS

- Important to be cautious when interpreting multiple analyses of same data set
- Some analyses other than primary analysis may be compelling
- Ever-increasing number of methods to account for multiple analyses
- Still best to rely on pre-specification (of hypotheses) and replication (of results)

QUESTIONS?