

# Multiple Comparisons

Brent R. Logan, PhD

Sponsored by the Clinical and Translational Science Institute (CTSI)  
and the Department of Population Health / Division of Biostatistics



# Speaker Disclosure

In accordance with the ACCME policy on speaker disclosure, the speaker and planners who are in a position to control the educational activity of this program were asked to disclose all relevant financial relationships with any commercial interest to the audience. The speaker and program planners have no relationships to disclose.

# Outline

- Introduction/Motivating Example
- Review of Hypothesis Testing
- Multiple Testing Strategies
- Philosophical Issues
- Specific Examples:
  - Multiple Outcomes/Timepoints
  - Multiple Groups
  - Subgroup Analysis in Clinical Trials
  - Large multiplicity problems
  - Interim Analyses

# Or, Torturing the Data until it Confesses



# Motivating Example

- Panic Disorder study to measure effectiveness of a drug treatment compared to control
- Four outcomes studied
  - Severity of anticipatory anxiety,  $p=0.04$
  - Total number of panic attacks,  $p=0.10$
  - Severity of phobic avoidance,  $p=0.72$
  - Global assessment of patient,  $p=0.38$
- Conclusion based on  $p < 0.05$  : Drug has a statistically significant effect on severity of anticipatory anxiety

# Multiplicity Problem

- Four outcomes were studied
  - Increased chance of finding at least one false significant finding the more tests you perform
  - Is the impact of treatment on severity of anticipatory anxiety “real” or is it an artifact of performing multiple tests?
- Extreme setting: Keep examining more and more outcomes until you find one with a significant p-value

# Review of Hypothesis Testing

- Null Hypothesis  $H_0$ :
  - Typically no difference between groups or no association
- Alternative Hypothesis  $H_1$ :
  - Research hypothesis, there is a difference or association between groups
- Compute p-value: Likelihood of obtaining an observed difference or one more extreme if the null hypothesis were true (i.e. by chance alone)
- Compare p-value to significance level  $\alpha$ : if  $p < \alpha$ , then reject  $H_0$

# Types of Errors

*Truth about Population*





$H_0$  is True

$H_0$  is NOT True

*Decision*

Accept  $H_0$

Reject  $H_0$

		
		
	Type I error	Type II error

$\alpha$  = probability of Type I error (level of significance)

$\beta$  = probability of Type II error

$1 - \beta$  = Power



# Testing Multiple Hypotheses

- K independent tests with type I error 0.05

K	Probability of at least one type I error	Expected number of type I errors
1	0.05	0.05
2	0.10	0.1
4	0.18	0.2
10	0.40	0.5
50	0.92	2.5
100	0.99	5
1000	1	50
10000	1	500





# Implications for our Example

- There is a 18% chance of obtaining a statistically significant result among our four endpoints even if there was no effect of treatment on the drug.
- $P=0.04$  seems less convincing given the context of multiple comparisons
- How do we incorporate the impact of multiple testing on our inference?

# Multiple Testing Strategies

- Perform less tests
- Transparency
- Cautious in interpretation
- Ad-hoc adjustment: Use significance level of 1% rather than 5%
- Control a different error rate which incorporates the number of tests performed
  - Familywise error rate (FWE): Probability of at least one type I error among all tests performed
  - False Discovery Rate (FDR): Expected proportion of “False discoveries” out of the total significant findings

# Error Rates

	True Null Hypotheses	False Null Hypotheses	Total
Decision	Accepted Hypotheses A 	B 	M
	Rejected Hypotheses C 	D 	N
	Total T	F	

$FWE = P(C \geq 1)$ ;  $FDR = \text{Average of } (C/N)$

# Controlling the FWE

- Overall test of whether there is an effect of treatment on all outcomes at once
  - Only examine individual outcomes if overall test is significant (weak control)
- Adjust significance level (equivalently, p-value) for number of tests performed ( $K$ )
  - Bonferroni procedure: Use significance level= $\alpha/K$ , or multiply all p-values by  $K$
  - Other, more efficient procedures available: See a statistician
  - Strong control

# Controlling the FDR

- Benjamini-Hochberg procedure
  - 5% FDR: On average, 5% of your significant findings will be false
  - Order the p-values
  - Compare  $i$ th smallest p-value to  $i \times 0.05/K$

# FWE vs. FDR

- Example comparing 2 groups on a psychological scale which has 14 different subscales
  - Comparison of groups for each subscale
  - 14 tests

# Benjamini-Hochberg

Ordered p-value

threshold

Bonferroni threshold

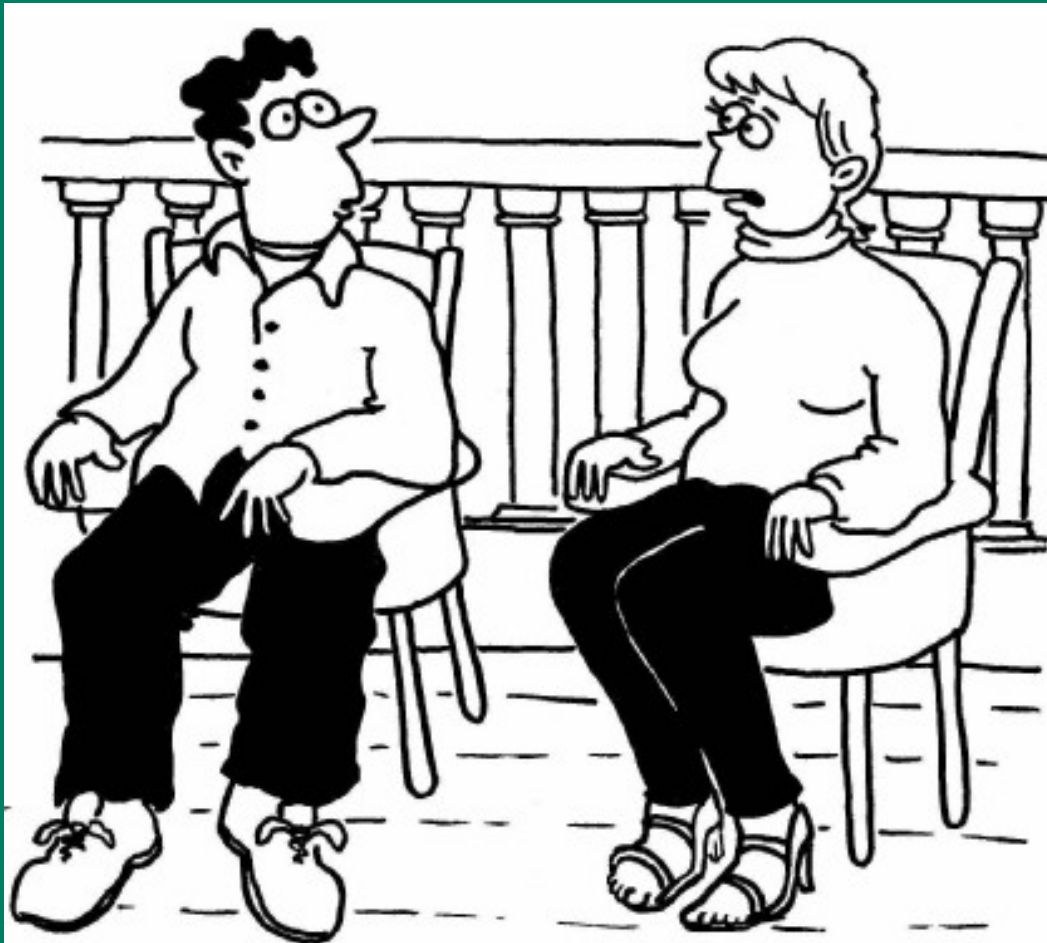
Ordered p-value	Benjamini-Hochberg threshold	Bonferroni threshold
0.00039	0.0036	0.0036
0.00103	0.0071	0.0036
0.00159	0.0107	<b>0.0036</b>
0.00164	<b>0.0143</b>	0.0036
<b>0.00765</b>	<b>0.0179</b>	0.0036
<b>0.0196</b>	<b>0.0214</b>	0.0036
<b>0.0237</b>	<b>0.0250</b>	0.0036
0.0310	0.0286	0.0036
0.101	0.0321	0.0036
0.157	0.0357	0.0036
0.284	0.0393	0.0036
0.542	0.0429	0.0036
0.543	0.0464	0.0036
0.793	0.0500	0.0036



# FWE vs. FDR

- FDR controlling procedures are generally more powerful than FWE controlling procedures
  - More likely to detect real differences as significant
  - Allows for an acceptable rate of type I errors among significant findings
  - Not appropriate when strict control of any type I errors is desired
  - Exploratory setting
  - Large scale multiple testing problems: Genetics, Imaging
- FWE controlling procedures:
  - Strict control of risk of any type I errors
  - More appropriate in confirmatory or regulatory setting

# What about confidence intervals?



**" When you say you're 95% confident . . . just what are you inferring ? "**

# Confidence Intervals

- Equivalence between hypothesis testing and confidence intervals
- Null Hypothesis: mean difference=0
  - Construct a confidence interval for the mean difference and see if it contains the null value of 0
  - Correspondence between type I error rate of hypothesis test (5%) and confidence level of interval (95%)

# Multiple Confidence Intervals

- 95% Confidence Interval: Probability the interval contains the true parameter is 95%
- 95% ***Simultaneous*** Confidence Interval: Probability that all intervals contain the true parameters is 95%
- Wider than individual CI's
- Direct correspondence to FWE
- If you adjust your p-values, you should also adjust your Confidence Intervals

# Philosophical issues

- Differing opinions on whether or when adjustment is needed
- Choice of family of hypotheses to adjust for is arbitrary, and results are very sensitive to this choice
  - Choose small, more focused families, specified a priori (in writing) to avoid cheating
- Increase in type II errors due to adjustment (loss of power or ability to detect real differences)
  - Use an *appropriate* and *powerful* testing procedure (see statistician)
- Argue for unadjusted analysis, but with full disclosure of data analysis procedures
  - Difficult for reviewers to evaluate

# Philosophical issues

- Need for adjustment, and best method of adjustment, is often scenario dependent
- Things to consider (Westfall et al., 1999)
  - Is it plausible that many of the null hypotheses might be true?
  - Do you want to ensure reproducibility, or be able to claim that an identified significant finding is in fact real?
  - Do you want to heavily mine the data to find a “significant” result?
  - Is your study expensive and unlikely to be repeated before serious actions are taken?
  - Is there an important cost associated with type I errors?

# Multiple Outcome Variables

- Limit number of outcomes
- If control of type I error is desired
  - Outcomes are often correlated
  - Bonferroni method is conservative
  - Overall test of no difference on any outcomes is more powerful (Multivariate methods)

# Multiple Outcome Variables

- Panic Disorder Motivating Example:  
Bonferroni adjustment

Outcome	P-value	Adjusted p-value
Severity of anticipatory anxiety	0.04	0.16
Total number of panic attacks	0.10	0.40
Severity of phobic avoidance	0.72	1.00
Global assessment of patient	0.38	1.00



# Multiple Outcome Variables

- Overall test:
  - Null hypothesis: No difference between treatment and control on any of the 4 endpoints
  - $p=0.20$
- No evidence of a difference on any of the endpoints
- Ignore significant p-value ( $p=0.04$ ) on severity of anticipatory anxiety
  - Likely attributable to multiple tests

# Multiple Outcome variables

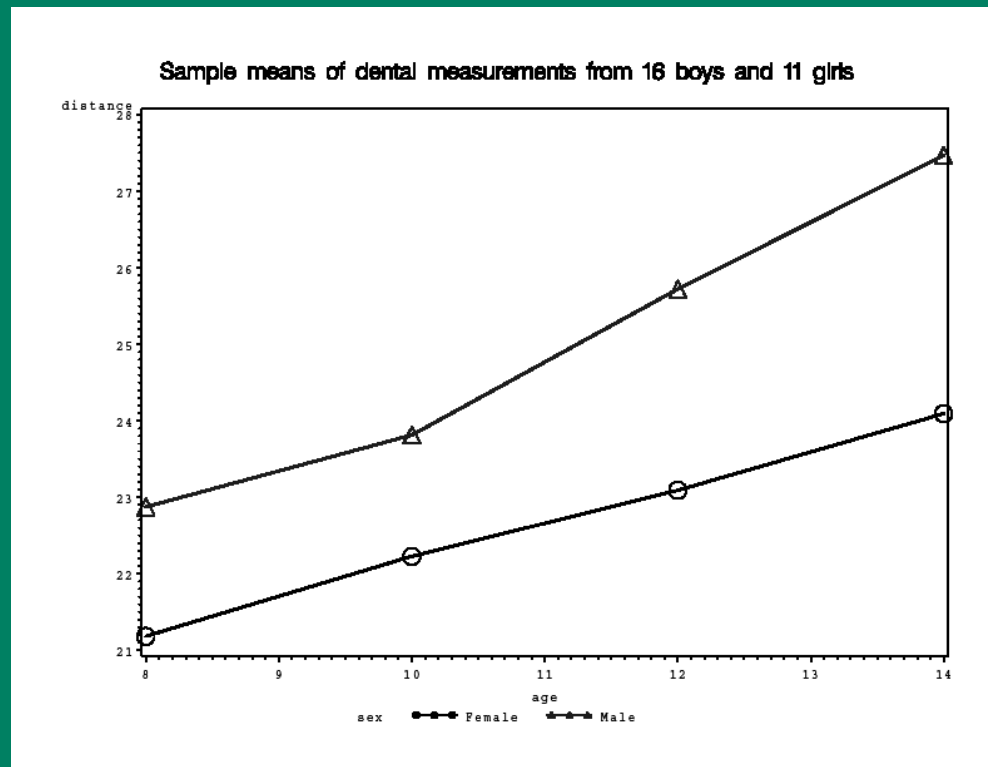
- Confirmatory clinical trials
  - Strict control of type I error rate desired
  - Primary endpoint:
    - usually one
    - pre-specified in protocol
  - Secondary endpoints:
    - Exploratory
    - Explanatory of findings in primary endpoint
  - No adjustment needed for primary endpoints

# Multiple Outcome variables

- Composite Endpoints:
  - Combine multiple endpoints into one to reduce multiple comparisons problem
  - Cardiovascular: Myocardial infarction, stroke, or cardiovascular death
  - Cancer: Progression-free survival
  - Potential difficulties in interpretation

# Multiple Time Points

- Repeated Measures study of dental measurements in boys and girls (Pothoff and Roy, 1964) at ages 8, 10, 12, 14



# Multiple Time Points

- Separate t-tests at each time point comparing boys and girls are susceptible to multiplicity issues
- Fit a linear growth curve
  - Reduces number of comparisons from 4 time points to 2 parameters (slope and intercept)
- Overall test
  - Null Hypothesis: No difference between boys and girls at any age
  - Comparisons at each age only performed if overall test is significant

# Multiple Time Points

- Comparisons between Boys and Girls at each age

Age	Mean difference (B-G)	p-value	Bonferroni Adjusted p-value
8	1.69	0.066	0.264
10	1.58	0.084	0.336
12	2.63	0.005	0.02
14	3.38	<0.001	0.002

# Multiple Time Points

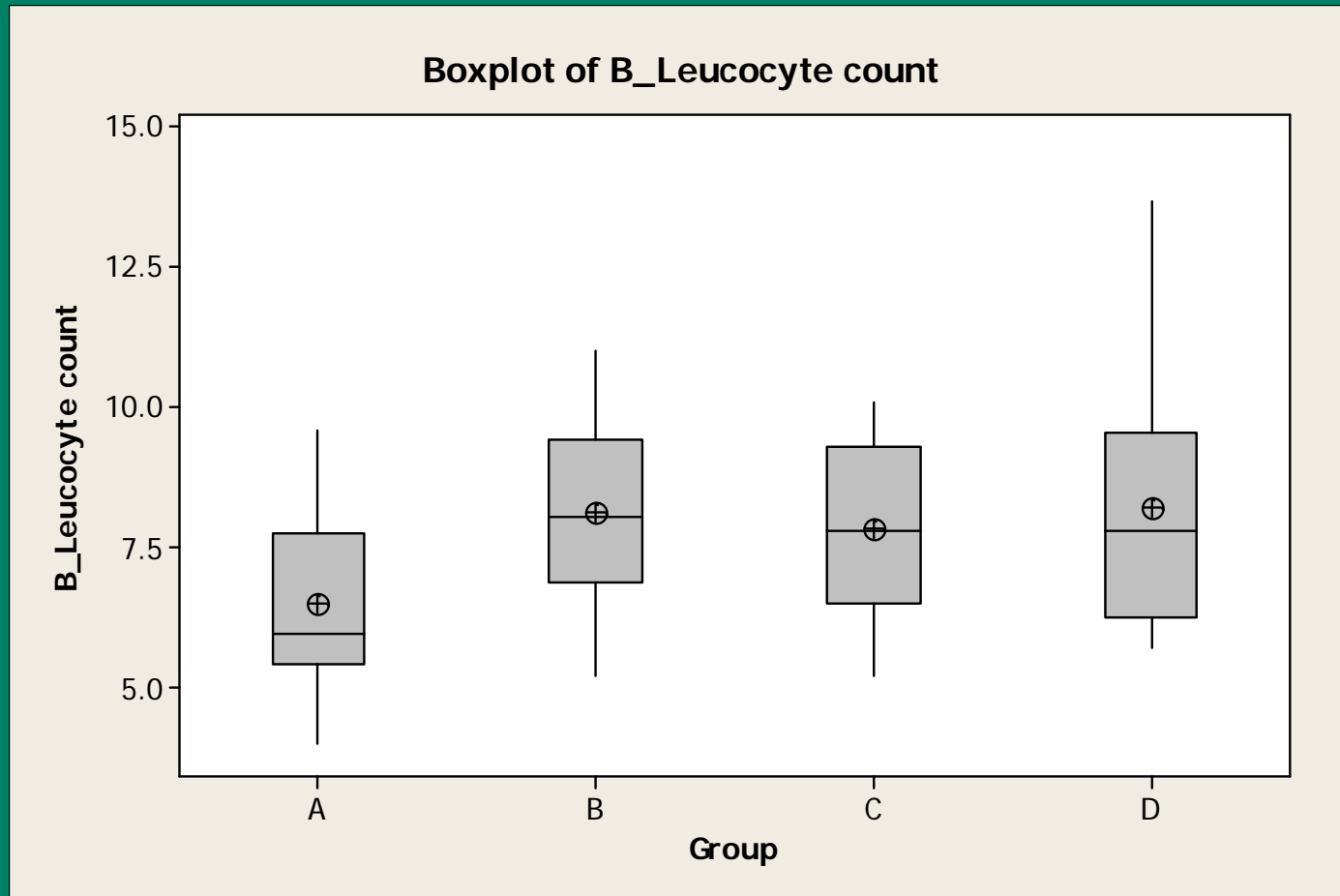
- Overall test:
  - Null Hypothesis: No differences between boys and girls at any age
  - $p=0.006$
  - Since this is significant, we look at unadjusted p-values for each time point
  - Significant differences between boys and girls at Age 12 ( $p=0.005$ ) and Age 14 ( $p<0.001$ )
- Both multiplicity adjustments give similar findings
- Final Conclusion: Mean dental measurements are significantly different between boys and girls at ages 12 and 14

# Multiple Groups

- Comparison of B-leucocyte counts in 51 subjects with colorectal cancer (Werther et al., 2002)
- Four classifications of cancer patients:
  - Duke's Classification A, B, or C
  - Group D=patients with disease which had not been completely resected
  - 6 pairwise comparisons



# Multiple Groups



# Multiple Groups

- Unadjusted p-values for pairwise comparisons (row vs. column)

	B	C	D
A	0.045	0.098	0.062
B		0.683	0.891
C			0.638

- Significant difference between Classification A and B ( $p=0.045$ )

# Multiple Groups

- Overall ANOVA F-test has  $p=0.191$
- Ignore significant result in pairwise comparisons when the overall test is not significant
- Final Conclusion:
  - No significant difference in B-leucocyte counts between cancer groups
- For strict control of type I error rate, standard method is Tukey test
  - More powerful than Bonferroni

# Multiple groups

- Adjusted p-values for pairwise comparisons using Tukey test

	B	C	D
A	0.183	0.342	0.237
B		0.976	0.999
C			0.964

- No significant differences among cancer groups, since all adjusted p-values are  $>0.05$ .

# Subgroup Analysis

- Is effect of treatment in a clinical trial homogeneous across all patients in that trial?
- Example 1: ISIS-2 Trial: 17000 patients with AMI randomized to placebo vs. aspirin (also streptokinase) (Lancet, 1988)
  - Mortality within 1 month: 9% (aspirin) vs. 12% (placebo),  $p < 0.001$
  - Investigators were urged (by editors) to conduct nearly 40 subgroup analyses
  - Investigators agreed on condition that they could conduct their own subgroup analysis to illustrate unreliability of subgroup findings

# Subgroup Analyses

- Subgroup defined by astrological sign

# of deaths in 1 month

Astrological sign	Aspirin	Placebo	p-value
Libra or Gemini	150	147	NS
Others	654	869	<0.001

- Increased variability of results just due to chance when you look at a lot of subgroups.
- Excess of type II errors due to multiple comparisons

# Subgroup Analysis

- Example 2: Effect of new vs. standard antibiotic on febrile morbidity in four age strata and overall

Age	Rate ratio	95% CI
20-24	1.4	(0.6-3.2)
25-29	1.2	(0.4-3.1)
30-34	0.3	(0.1-0.9)
35-39	1.1	(0.5-2.5)
Overall	0.9	(0.6-1.4)

- Analysis in 4 subgroups inflates the type I error rate

# Subgroup Analysis

- Proper assessment of subgroups
  - Perform overall test of whether the subgroups differ
  - Null Hypothesis: treatment effect is the same for each subgroup (No interaction between subgroup and treatment)
  - Only look at each subgroup if interaction test is significant
- Example 2: Interaction test:  $p=0.103$ , no evidence of interaction
  - Subgroup finding in age 30-34 group is likely due to chance



# Subgroup Analysis

- Subgroup analyses are discouraged and prone to overinterpretation
- Recommendations
  - Perform interaction test: only look at each subgroup separately if interaction test is significant
  - Confine to primary outcome and limited number of subgroups
  - Prespecified in protocol
  - Consider biological plausibility
  - Report all subgroup analyses done – may need to adjust for multiple subgroup variables (age, sex, disease status)
  - Generally considered ***EXPLORATORY***

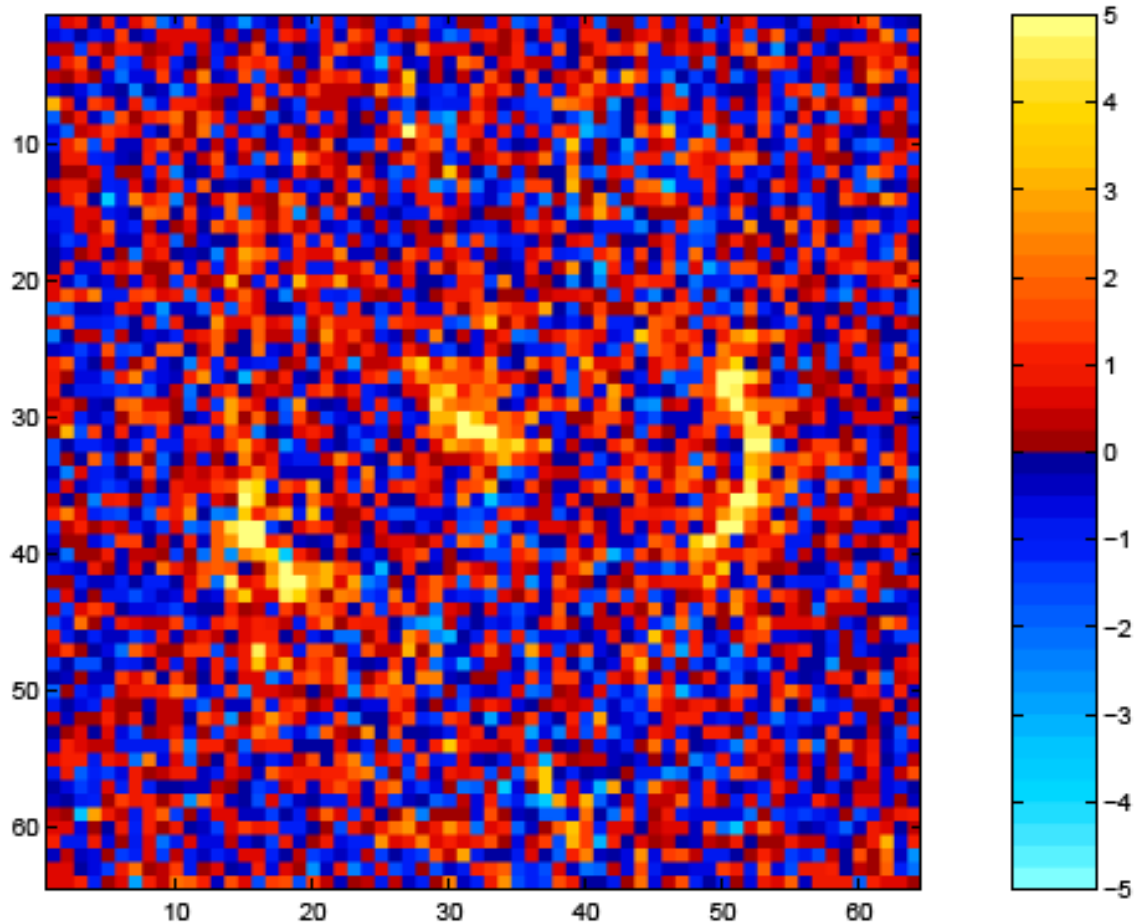
# Large Multiplicity Problems

- Genetics and microarrays
  - Thousands of genes assessed for their association with phenotype
  - Hypothesis test performed for each gene
  - If each test is performed at the 5% significance level, we would expect 50 false findings in 1000 tests.

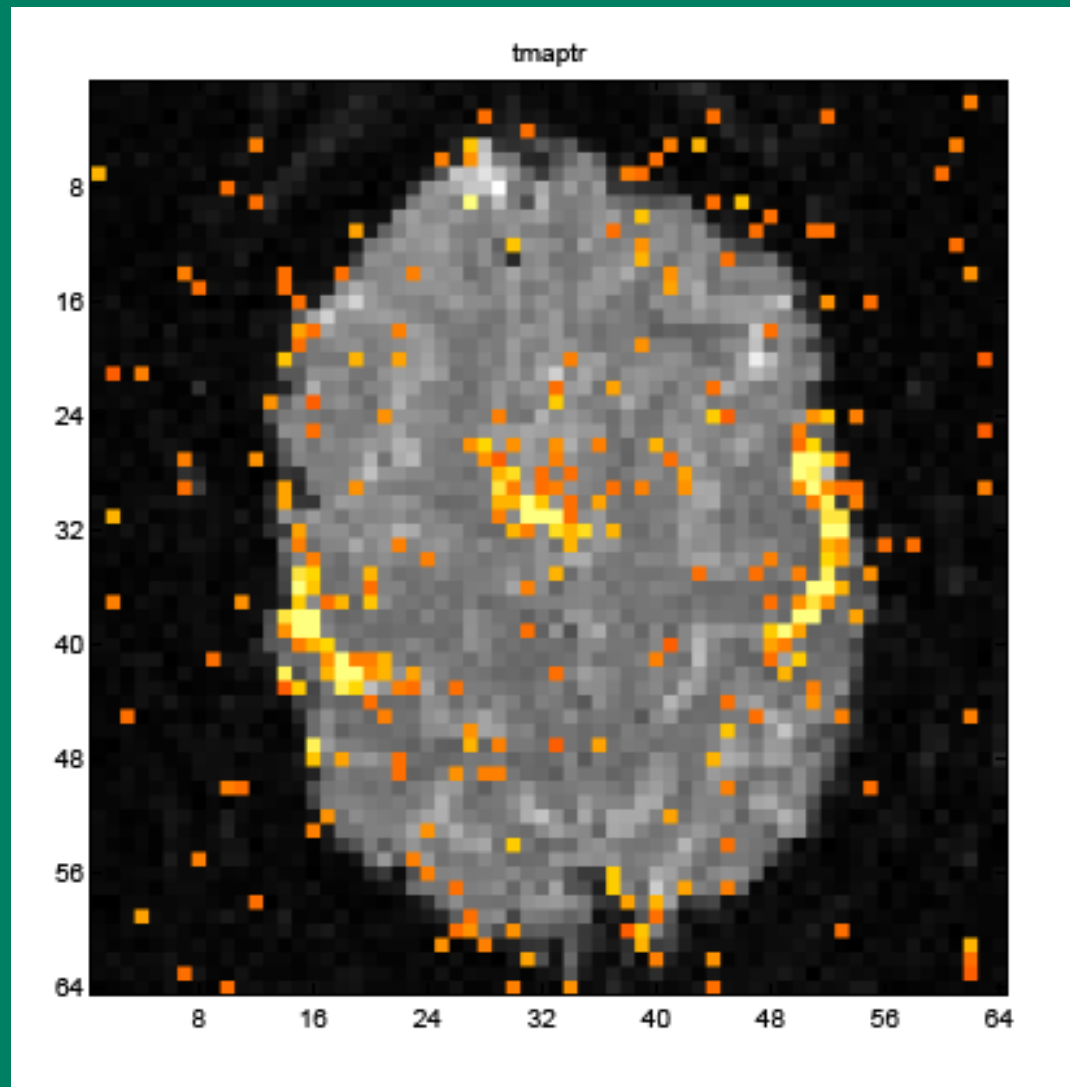
# Large Multiplicity Problems

- Functional Magnetic Resonance Imaging
  - Experiment performed and Blood Oxygen Level Dependent (BOLD) response assessed at each of thousands of voxels or points in the brain
  - Example: Finger-tapping experiment
  - Null hypothesis: Mean BOLD signal while tapping is the same as the Mean BOLD signal while not tapping
  - T-test performed at each point in the brain

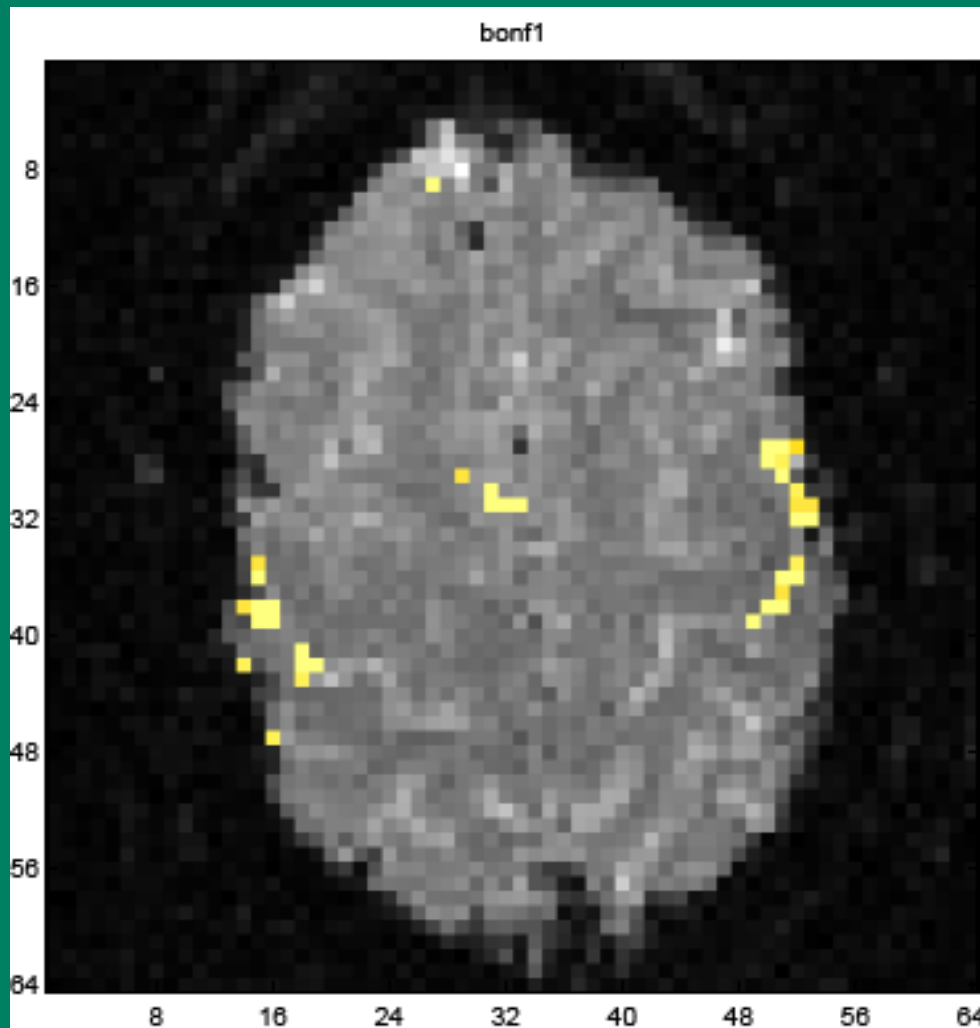
# T statistic image



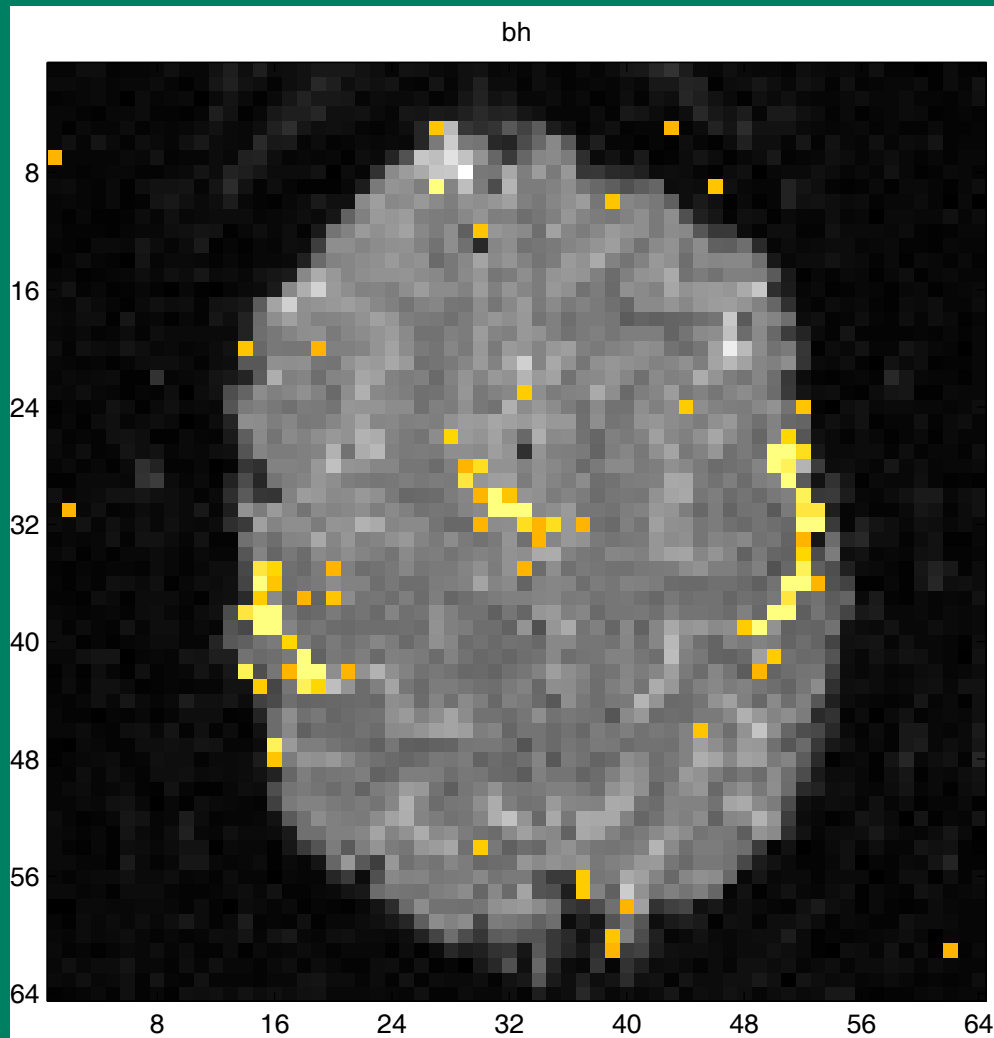
# Thresholded at 1.96



# 5% FWE Threshold (Bonferroni)



# 5% FDR Threshold



# Large Multiplicity Problems

- Unadjusted method: Too many false positives
- FWE adjustment: Loss of power to detect real effects
- FDR procedures: Compromise
  - Low rate of false positives relative to true discoveries
  - Improved power relative to FWE adjustment



# Interim Analyses

- Clinical Trials are routinely monitored for safety by a DSMB
  - Meets every 6-12 months to review data
- Justifications for early termination of study
  - Unacceptable toxicity
  - Accrual problems
  - Efficacy
  - Futility

# Interim Analyses

- Repeated tests for efficacy can inflate the type I error rate

# of looks	Type I error rate
1	0.05
2	0.08
3	0.11
10	0.19

# Interim Analyses

- Plan must be specified in the protocol
- Adjust significance level at each interim analysis
- For example, with 3 looks at the data

Interim analysis	Adjusted significance level (O'Brien-Fleming)
1	0.0005
2	0.014
3 (final)	0.045

# Summary

- The more tests you do the more likely you are to find a significant result
- Restrict or prioritize the number of tests
- Be cautious and aware when interpreting results
- Explicit corrections for multiple testing are available
  - Strengthen evidence for significant research findings
  - Loss of power
- Need an appropriate
  - Family of hypotheses
  - Type of error rate

# Resources

- The **Clinical and Translation Science Institute** (CTSI) supports education, collaboration, and research in clinical and translational science: [www.ctsi.mcw.edu](http://www.ctsi.mcw.edu)
- The **Biostatistics Consulting Service** provides comprehensive statistical support <http://www.mcw.edu/biostatsconsult.htm>

# Free drop-in consulting

- **MCW/Froedtert/CHW:**
  - Monday, Wednesday, Friday 1 – 3 PM @ CTSI Administrative offices (LL772A)
  - Tuesday, Thursday 1 – 3 PM @ Health Research Center, H2400
- **VA:** 1<sup>st</sup> and 3<sup>rd</sup> Monday, 8:30-11:30 am
  - VA Medical Center, Building 70, Room D-21
- **Marquette:** 2<sup>nd</sup> and 4<sup>th</sup> Monday, 8:30-11:30 am
  - Olin Engineering Building, Room 338D

# Upcoming Lectures

<p><b><u>August 14</u> at 7AM (Aniko Szabo, PhD)</b> Concepts on the Way from Data to Decisions Location: NT22009</p>	<p><b><u>September 24</u> at 1:30PM (Brent Logan, PhD)</b> Designing Clinical Trials Location: CHW Auditorium</p>
<p><b><u>August 26</u> at 8:50 AM (Prakash Laud, PhD)</b> Concepts on the Way from Data to Decisions Location: TBA</p>	<p><b><u>September 30</u> at 8:50 AM (Jennifer Le-Rademacher, PhD)</b> Statistics, Probability and Diagnostic Medicine Location: TBA</p>

For locations that are TBA please check the website below two weeks prior to the lecture date:

<http://www.mcw.edu/biostatistics/CalendarCurrentEvents/SeminarSeries.htm>