# Ethical Considerations in Study Design and Analysis

Linda Valeri
Laboratory for Psychiatric Biostatistics
McLean Hospital

April 27, 2016

# Outline

- Review some features of good study design, importance of statistical plans and transparent reporting and their relevance to ethical research.

- In context of clinical trials, some excellent guidelines have been developed, e.g., CONSORT guidelines on complete and transparent reporting of trials.
  http://www.consort-statement.org

- Although main focus is on aspects of study design and analysis of intervention studies, these have implications for observational studies as well.

# Ethics in Research

The integrity of scientific research can be undermined in a number of ways.

Unethical behavior:

- Intentional misconduct
- Intentional fraud

More commonly, due to misleading findings arising from studies with poor designs and analysis plans and less than transparent reporting of results.

# Fraud and Misconduct


**MMR vaccine link to autism**

# Fraud and Misconduct

In 1998, paper published in Lancet by Wakefield and 12 co-authors claiming link between MMR vaccine and autism.

Wakefield et al reported on 12 children who developed symptoms of autism within 2 weeks of MMR vaccination.

Lancet paper fueled an MMR scare that quickly took off in U.K. and soon after around the world.

12 years later, paper was retracted from Lancet.

# Fraud and Misconduct

In 2011, BMJ paper by Deer exposed bogus nature of data:

- 3 children reported with "regressive autism" did not have autism at all.
- 5 children had documented pre-existing developmental concerns prior to MMR vaccination.
- Wakefield altered or misrepresented medical records of all 12 children.

# Fraud and Misconduct

Public health consequences:

- Immunization rates in Britain dropped from 92 percent to 73 percent.
- 2014 saw the highest measles case count in U.S. (667) since the disease was declared eradicated in the U.S.
- In 2015, 189 cases of measles were reported in the U.S. (Source: CDC)

# Ethical Research

Study designs, statistical plans and transparent reporting are key components of ethical research:

(i) Features of study design
(ii) Sample size considerations
(iii) Data analysis/statistical methods
(iv) Valid interpretation of the findings
(v) Transparent reporting in scientific publications

# Study Design

Fundamentals:

Is the research question well-formulated?

Does the study design address the main research question (hypothesis)?

Is the outcome variable clearly defined (prior to obtaining preliminary results)?

# Study Design

Importance of Control Group

- Is there a control group?
- A group comparable to the intervention/exposure group in every way except the intervention/exposure.

  Note: control group may be composed of subjects receiving no intervention, a different intervention, or the same intervention but administered at different schedule/dose.

# Study Design

## Randomization

Three major advantages to randomization:

(i) Selection bias is eliminated from assignment of interventions. Comparisons not invalidated by selection of patients of a particular kind, consciously or not, to receive a particular form of intervention.

(ii) Tends to balance intervention groups in prognostic factors, whether or not these variables are known.

(iii) Guarantees validity of statistical tests of significance.

# Study Design

Type of Randomization:

(i) **Simple** randomization: determine each patient's intervention at random independently with no constraints.

(ii) **Block** randomization: the experimenter divides subjects into subgroups called blocks, such that the variability within blocks is less than the variability between blocks and to equalize the number of subjects on each treatment.

(iii) **Stratified** randomization: Achieve approximate balance on important prognostic characteristics, e.g., disease severity.

# Study Design

**Blinding**:

Prevents the blinded parties from intentionally or unintentionally affecting the results through their knowledge of the intervention status.
Blinding of groups at potentially many levels:

- Subjects involved in study
- Investigators
- Data collectors
- Outcome adjudicators
- Data analysts

# Study Design

Sample Size Considerations:

- What is the justification of the study sample size?
- From ethical perspective, if too few subjects, investigator cannot adequately address the study question.
- If more subjects enrolled than needed, then too many subjects unnecessarily exposed to potential risk.

# Statistical Analysis Plan

Is there a complete data analysis plan?

Avoids risk that investigators choose the analysis based on the results obtained, which would invalidate statistical assessment.

Data analyses should be

- Consistent with original data analysis plan
- Clearly described
- Reproducible by someone else if you provide the data

# Some Common Pitfalls

Problems of "multiplicity" are very common.

Pre-specified versus post-hoc analyses (recall earlier comments about importance of statistical analysis plan).

General lack of transparency is another major concern: Many ethical dilemmas arise from how study results are reported.

# Problems of Multiplicity

Problems of multiplicity commonly arise when:

(i) a small number of groups (e.g., treatment and control) are compared in terms of many outcomes, or

(ii) many sub-groups (e.g., defined by various baseline characteristics) are compared in terms of a single outcome.

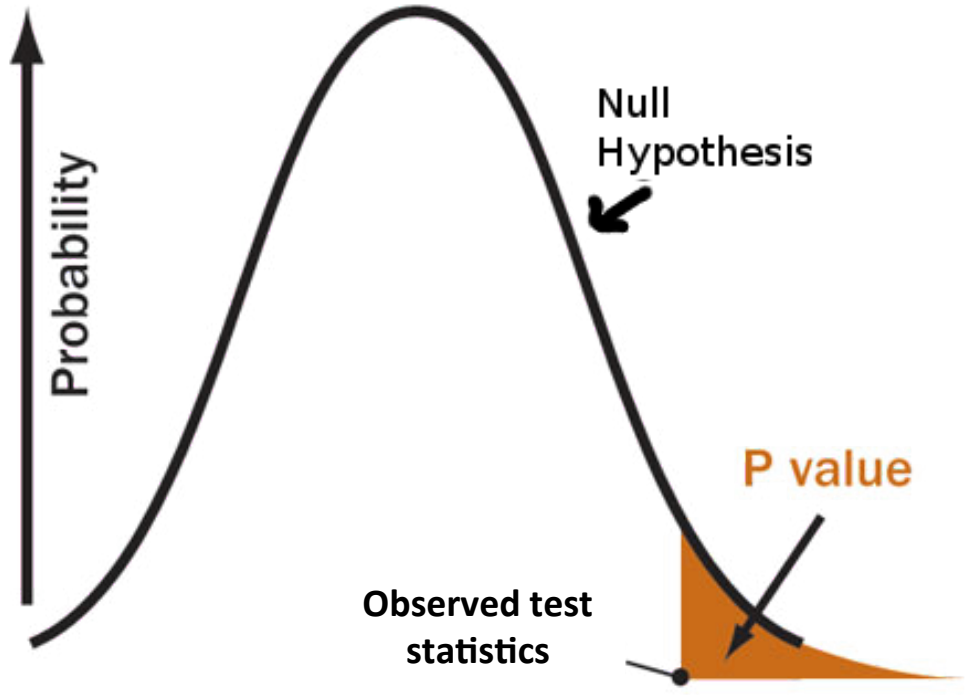
Both scenarios are problematic because multiple tests inflate so-called "type I errors".

# Hypothesis Testing

**Recall:** Hypothesis testing provides a framework for making inferences based on the observed outcomes of an experiment/study.

A null hypothesis, $H_0$, is the hypothesis of no effect (e.g., no difference between groups).

An alternative hypothesis, $H_A$, is a hypothesis about an effect the researcher would like to establish.


Probability
Null
Hypothesis
P value
Observed test
statistics

# Type I & Type II Errors

A statistical test can yield two types of errors:

type I error is the error of rejecting $H_0$ when it is true

type II error is error of not rejecting $H_0$ when it is false

The **level** of the test is the maximum probability of a type I error under $H_0$ (by convention, typically set at 0.05).

The **power** of the test, at a specific alternative, is the probability of correctly rejecting $H_0$ (typically 0.80-0.90) (recall earlier comments about sample size)

# P-value

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

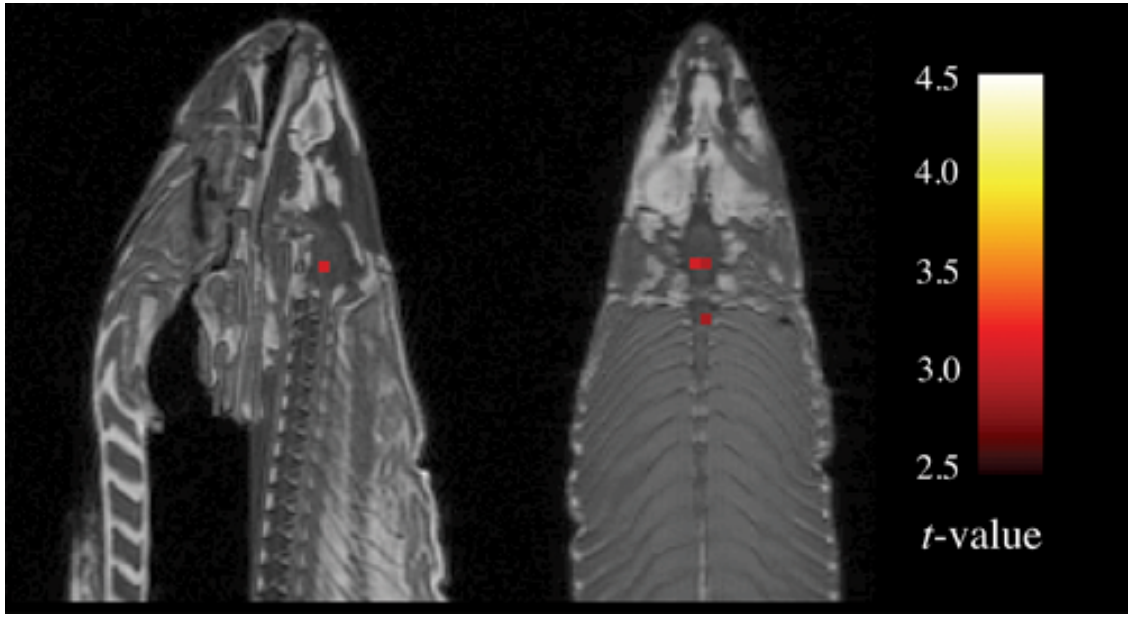
*A p-value of 0.05 signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed.*

# Multiplicity inflates Type I Errors

When many tests conducted, each at 0.05 level, the probability of a type I error can be greatly inflated.

For example, when 12 (independent) tests are conducted the chance of a type I error is 46% (or $1 - (1 - .05)^{12} = 0.46$).

# Why is multiplicity a problem: Life after death?



# Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

## INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

## METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the fMRI timeseries, coregistration of the data to a $T_1$-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.

Analysis. Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was applied.

Voxel Selection. Two methods were used for the correction of multiple comparisons in the fMRI results. The first method controlled the overall false discovery rate (FDR) and was based on a method defined by Benjamini and Hochberg (1995). The second method controlled the overall familywise error rate (FWER) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).

## DISCUSSION

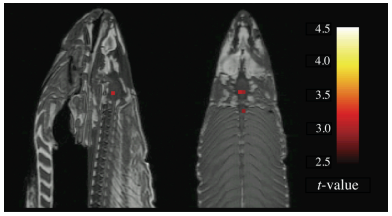
Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

## REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.

Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

## GLM RESULTS

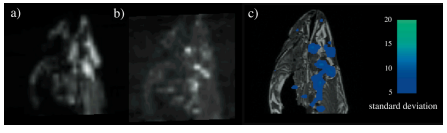



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 $mm^3$ with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.

Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

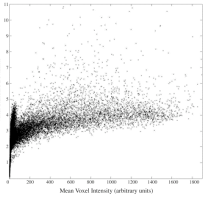
## VOXELWISE VARIABILITY




To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.

We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution $T_1$-weighted image.

To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.



# Pre-Specified versus Post-Hoc Tests

Pre-specified: planned prior to examining the data.

Post-hoc: not specified prior to examining the data.

However, in either case, both are subject to problems of multiplicity.

# Corrections for Multiplicity

**Formal:** Apply stricter criterion for judging statistical significance.

Bonferroni: If conducting 10 tests then use a criterion of $0.05/10 = 0.005$ to ensure no greater than 5% chance of type I error.
Note: $1 - (1 - 0.005)^{10} = 0.049$

**Downside:** Can be conservative, especially when tests are dependent or correlated.

# Corrections for Multiplicity

**Informal:** Acknowledge the number of nominally significant tests that would be expected to occur by chance alone.

For example, if conducting 40 tests at 0.05 level then note that 2 significant tests are expected to occur by chance alone.

This consideration can then be incorporated in the interpretation of the results.

# ASA Statement on p-values

In February, 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an American Statistical Association (ASA) discussion forum:

> Q: Why do so many colleges and grad schools teach p = .05?
>
> A: Because that's still what the scientific community and journal editors use.
>
> Q: Why do so many people still use p = 0.05?
>
> A: Because that's what they were taught in college or grad school.

# ASA Statement on p-values

The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions.

Mis-use of p-values is blamed for much of these issues.

Skepticism lead to radical choices, such as the one taken by the editors of Basic and Applied Social Psychology, who banned p-values (Trafimow and Marks, 2015).

In response to the recent "reproducibility crisis" ASA Board pronounced a statement on p-values and statistical significance published on February 2016.

First time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics.

# ASA Statement on p-values

## Principles

(i) P-values can indicate how incompatible the data are with a specified statistical model.

(ii) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by chance.

(iii) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

(iv) Proper inference requires full reporting and transparency.

(v) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

(vi) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

# Transparency/Completeness in Reporting

To properly evaluate results of a study, reviewers and readers must have adequate information about

(i) study design,

(ii) data collection,

(iii) data analysis, and

(iv) research findings

# Transparency/Completeness in Reporting

**Randomization**

- Was there randomization of interventions?
- What was the method of randomization?
- Haphazard assignment to groups is **not** random

**Blinding**

- Was the study blinded by design?
- Was the study blinded in reality?
- Example: Psychiatric diagnosis can sometimes be deduced from behavior

# Transparency/Completeness in Reporting

**Multiple comparisons**

- Were any significant findings discovered as part of a larger set of significance tests?

- If so, are the findings reported in the context of the larger set of tests, and a multiplicity correction applied or the limitations acknowledged?

**Missing data and Outliers**

- Missing data should be reported along with approach used to deal with it (e.g. complete case, imputation..)

- Were any outliers excluded from data analysis?

- If so, what were the criteria for their classification?

# Transparency/Completeness in Reporting

**Description of Analyses**

- Are you extensively describing modeling assumptions?
- Have you considered adding code in the supplementary material if the analyses are not trivially reproducible?

**Reporting of the findings**

- No "p-hacking"
- Report point estimates and measures of uncertainty (standard errors and confidence intervals)

# Transparency/Completeness in Reporting

Published papers should provide clear description of how study was conducted and what was found.

Similar to COI declarations, some journals are now requiring a "transparency declaration":

*"The lead author* affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained." (BMJ)*

Subsequent revelation of withheld or incorrect information is evidence of scientific misconduct.

# Tips for Reproducible Research

How to make data analysis consistent with original analysis plan:

Have an analysis plan!

Make adjustments before obtaining preliminary results whenever possible

Identify any post-hoc additions to your analysis in your methods and results

# Tips for Reproducible Research

How to make data analysis reproducible:

Use script for data analysis rather than relying on series of steps that must be reconstructed, e.g. from pull-down menus

If you use simulation-based approaches, set the seed

Archive a dataset, data analysis script, and output for each published paper

Some journals already require use of central data repositories for material used to prepare and support a publication

# Summary

A basic understanding of statistics has become a requirement for consuming/producing research in psychiatry.

Many opportunities in design of a study and data analysis to minimize bias and inflate type I error (e.g. randomization, blinding, multiple comparisons, treatment of outliers).

# Summary

Careful attention to study design and all steps of data analysis process (planning and reporting) can help address some of these problems.

Greater transparency allows reviewers/readers to judge a study's reliability and relevance.

Benefits: reproducible results more likely to positively impact science.

# Acknowledgements

Garrett Fitzmaurice, Director of Psychiatric Biostatistics Laboratory

Thank You!

Questions?

lvaleri@mclean.harvard.edu