Multiplicity – Incentives and risks of one fooling oneself

Michelle Detry

August 22, 2018

Disclosures & Statements

- Financial Disclosures:
  - Employee of Berry Consultants (Multiple Clients)
- Off-label statement
  - This presentation will not include information on unlabeled use of any commercial products or investigational use that is not yet approved for any purpose.

Multiplicities – the concern

- Multiplicities/Multiple Comparison concerns arise when numerous statistical tests are performed
- Some may not be “real”, may be “significant” due to chance alone
- Type I error – probability of making an incorrect conclusion of an effect
- The more tests you make the greater chance of making a Type I error

How much of a concern?

- Research question/goal of project determines amount of error willing to risk
- For example: Phase 3 confirmatory clinical trial
  - High level of evidence, want to change clinical practice
  - More stringent control of Type I errors

How much of a concern?

- Exploratory study, earlier phase study
- Maybe want to generate hypotheses
- Maybe less stringent control of Type I error because will have subsequent study designed to specifically to confirm results
- Type I error level should be set in context of your study question
- Also think what will you do next

Consequence of Lots of Data

- You get your study funded
- Want to get your money/effort's worth
- Collect lots and lots of information
- Easy to do with computing resources
- Can process lots of data
- Very very tempting to explore anything and everything to "find something significant"
- But you want credible and reproducible results too

Multiplicity Examples

- Multiple tests across multiple treatment arms
- Multiple tests among subgroups
- Multiple statistical tests performed while data is accumulating

Multiple Treatment Arms

- Randomized trial with 4 active doses of a new drug and a control/placebo arm
- Want to compare each active dose to control
- Four tests
- Let's say doses are ordered, 1, 2, 3, and 4 mg
- You run the trial and find p-values of 0.20, 0.18, 0.04, 0.30
- Do you believe that dose 3 is only effective dose?
- Will others believe it?

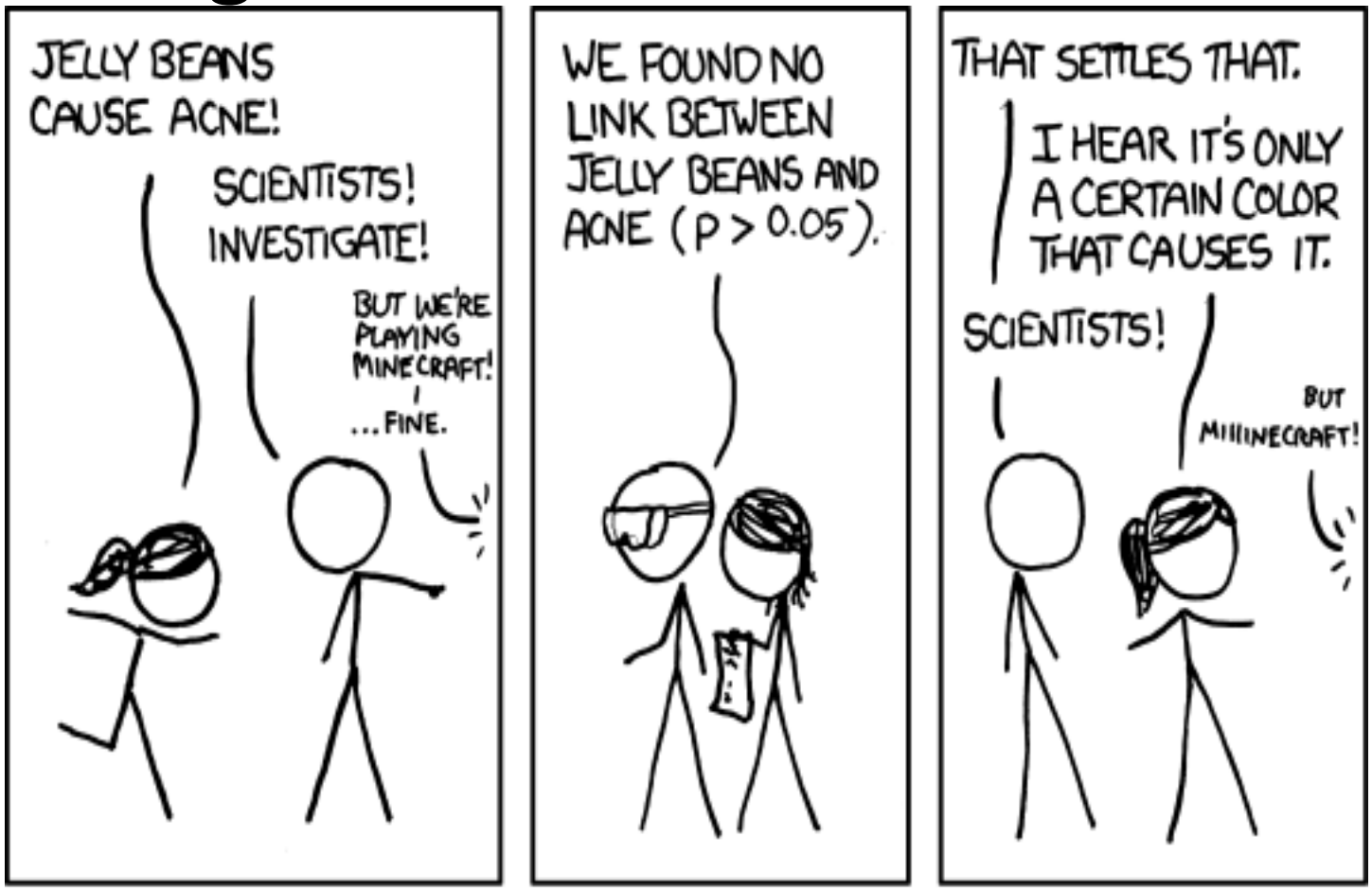
Multiple Treatment Arms (continued)

- Must think in the context of the question and background
  - What was research question
  - What was study designed to answer
- With these 4 doses we were thinking there would be a dose response, i.e. response would increase as dose increases
- Simply comparing each one to control did not really address the question
- Don't blindly live by p-values

Multiple Treatment Arms (continued)

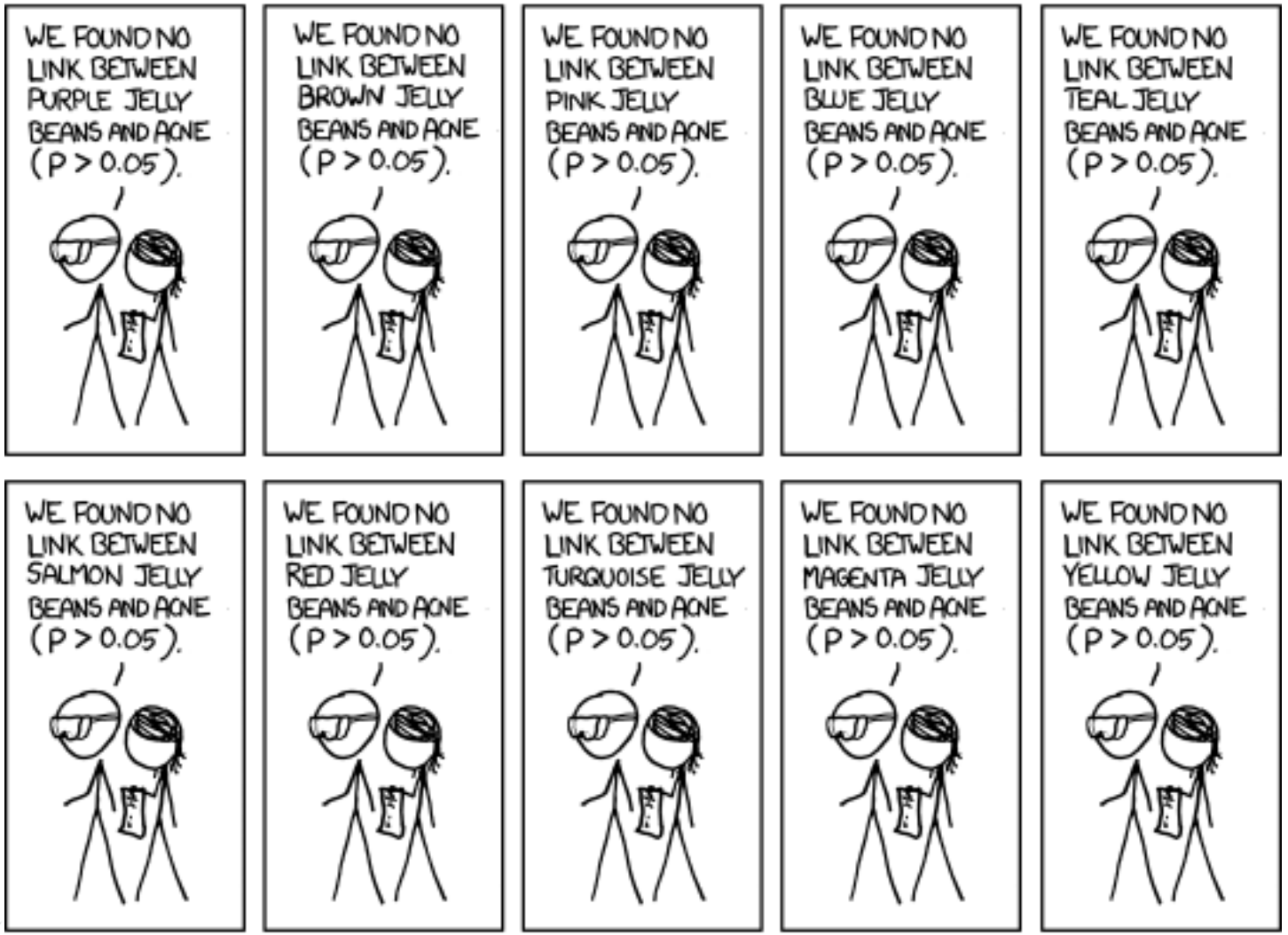
- Think about these issues and study question while designing trial and writing protocol
- Would have pre-specified how I would pick the best dose(s) and modeled the dose-response
- I would have thought, what was next – likely a confirmatory trial comparing the best dose(s) to the std of care/control

Multiple Subgroups: "Significant" xkcd.com comic



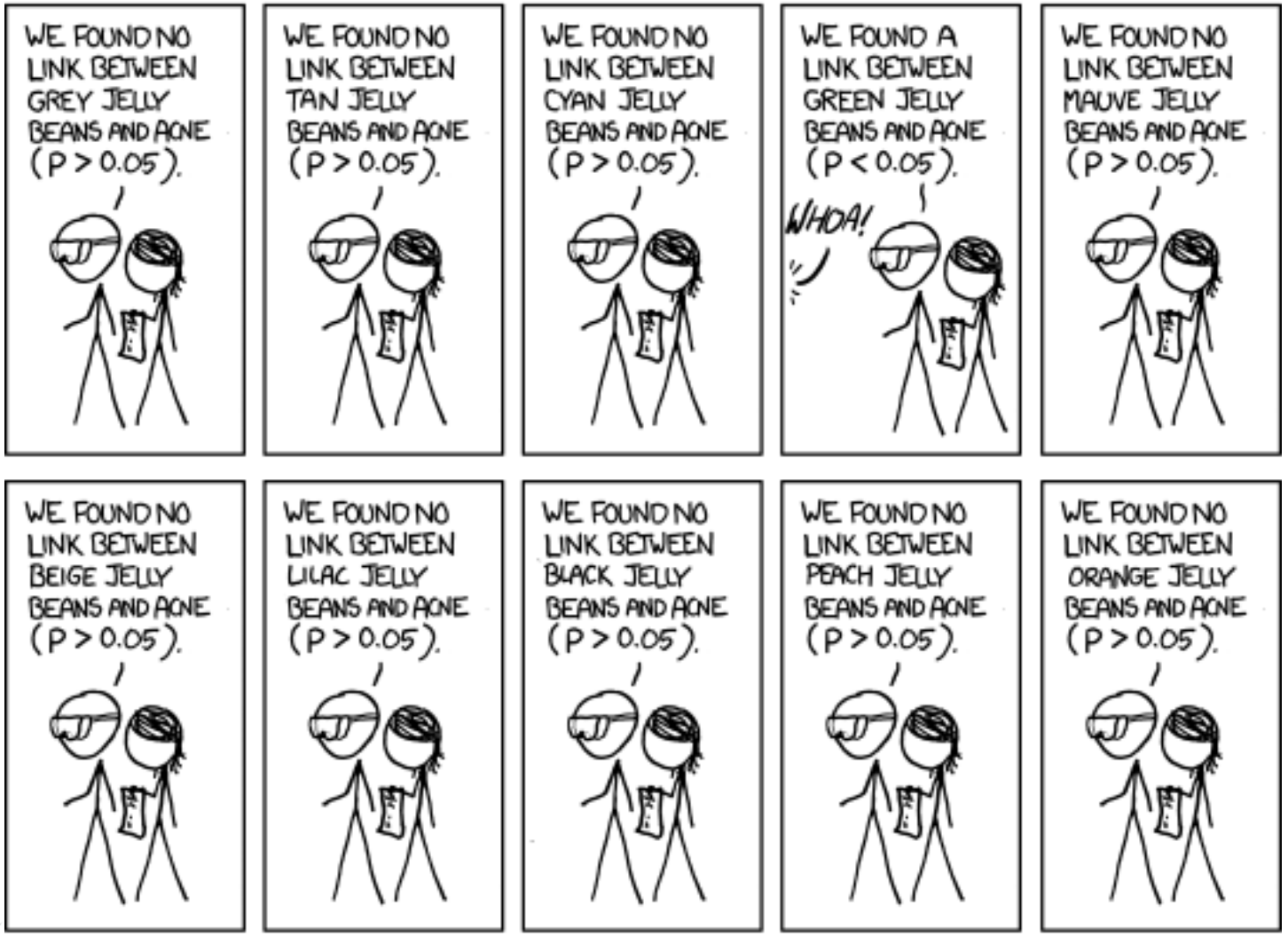
Berry Consultants
Statistical Innovation

“Significant” xkcd.com comic



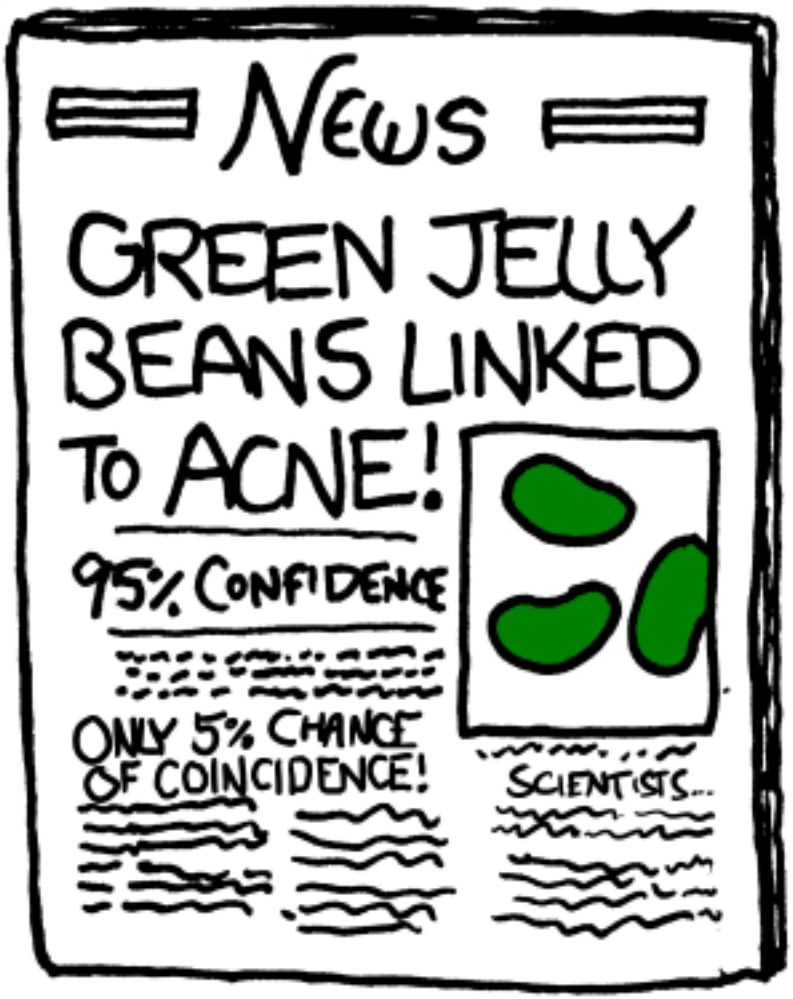
Berry Consultants
Statistical Innovation

“Significant” xkcd.com comic



Berry Consultants
Statistical Innovation

“Significant” xkcd.com comic



Berry Consultants
Statistical Innovation

Multiple Subgroups

- Randomized study comparing intervention A to standard of care
- At end of trial, unexpectedly, did not find significant effect
- But have lots of data
- Look at outcome in male, age>50, subjects with pre-existing disease X
- You see a benefit with intervention A!

Multiple Subgroups (continued)

- Do you believe it with the evidence in this trial?
- Would you have restricted inclusion to men, older than 50, with pre-existing disease X
- You want to publish result
- How should you frame the result in the context of multiple subgroup combinations you examined?
- Be honest, state what you did

Multiple Testing

- Example 1
- Your trial has half the subjects enrolled and data/outcomes collected
- You did not plan ahead that you would do an interim analysis of data
- But an abstract deadline in 2 weeks and you really want to submit
- Do you analyze the data and submit an abstract if you see a positive result?
- And then you will update the results with the full data when you go to the conference?

Multiple Testing (continued)

- Example 2
- You have a trial with a continuous outcome measure
- You measure a blood level, and want to see if there is a difference in outcome between good and bad blood levels
- What is the cutpoint for the blood level
- You try 1 definition, not significant
- You try 2nd definition, not significant
- You try a 3rd definition, significant!
- Do you report that there is a difference using the 3rd definition?

Berry Consultants
Statistical Innovation

Multiple Testing (continued)

- Example 3
- Larger double-blinded randomized clinical trial, that takes 3 years to enroll

- Two arms: intervention/control
- Outcome measured at 3 months

- Want to “look” at the data (unblinded) during trial to determine if you can stop early for success

Multiple Testing (continued)

- Acceptable methods to incorporate interim analyses
- Requires pre-planning (but that is always good) and pre-specification of design details
- Requires a higher level of evidence at earlier interims than later interims

Example

- Clinical trial for TBI, examines experimental treatment versus standard of care (SOC)
- Not sure how fast treatment needs to be administered, think sooner is better
  - Within 2 hours after injury
  - 2-12 hours after injury
- Want to compare active treatment to SOC in both populations

Example (continued)

- Want Type I error at 5%
- But we have 2 tests, increased chance of making Type I error
- Do I have to adjust for the multiple tests?
- Bonferroni adjustment to adjust for multiplicities would use 2.5% alpha for each test
- But with 2.5% alpha power is lower
- Need to enroll more subjects in each group to get back to desired power

Example (continued)

- Think about the question
- Would you think that giving the intervention quicker would yield better results?
- So if the intervention didn't work when given 0-2 hrs. after injury, would you think longer than 2 hrs. would work?
- So maybe instead of harsh Bonferroni adjustment, use a different adjustment method

Example (continued)

- One option:
  - Test the active trt vs. control in the 0-2 hrs. group
  - If it is significant at 0.05, then and only then do you test in the 2-12 hrs. group
  - If 0-2 hrs. population not significant, do NOT test 2-12 hrs. population
  - Pre-specify in protocol
  - Preserves power for comparison in the 0-2 hr. group
  - Can be shown that it preserves the Type I error rate

How to handle multiplicities

- Recommend much thought into defining study question
- Select primary aim and primary outcome
- Think about what you want to be able to say at the end of the trial
- Limit this question to single goal if possible
- Can have other aims, secondary and exploratory that do not require same level of evidence, conclusions are more exploratory

How to handle multiplicities

- Pre-specification!
- Define primary and secondary analyses in protocol
- Describe how you will handle multiple tests or defend if no adjustments will be made

How to handle multiplicities

- When reporting results, be clear on how you arrived at results
- Was it the primary analysis? Exploratory subgroup analyses?
- How many subgroups did you examine?
- Were these subgroups pre-defined?
- Interpret correctly and do not overstate

Adjustment for multiplicity

- Different statistical methods of adjustment for multiple tests
- Usually requires more evidence, a higher probability result is real
- However, consequence is that you may make more Type II errors where you conclude there is not an effect but there was
- Need to balance desire for multiple tests with consequences

Summary

- Multiplicities – it's complicated
- In publications/summaries be clear what was found through pre-specified analyses what wasn't
- Need to keep track of everything you analyze!
- Adjust where warranted, but better to focus during design phase to minimize need to adjust statistically

Summary

- Want your results to be reproducible! If results from exploratory analyses may need another study to confirm
- Consider results carefully in context of research question, prospective biological rationale, previous published studies
- If prepare multiple publications be clear about all the analyses, publications planned