Controversies and Unresolved Issues in the Design of Randomized Controlled Trials Testing Clinical/Behavioral Public Health Interventions

Part II: Rejecting Universal Adjustment for Multiple Testing in Public Health RCTs of Clinical/Behavioral Interventions

UCSF CAPS Methods Core Seminar

October 23, 2018

Steve Gregorich

Background

Ongoing debate about whether 'null' hypotheses and significance tests (use of *p*-values) are useful constructions.

That debate ebbs and flows
I am not going to dive into that

I believe that most journal editors and reviewers expect *p*-values to be reported when summarizing the results of RCTs

So, I am here talking to you about *p*-values

What is the 'multiple testing problem'?

α is the probability of making at Type I error
(falsely rejecting the null hypothesis; Neyman-Pearson).

Say you perform k=20 statistical tests, each at α=0.05.

If you assume…

. The null hypothesis is *true* for each test, *and*

. Each test is *independent*

Then you would expect

. $k \times \alpha = 20 \times 0.05 = 1$ test to result in a p-value <0.05 by chance

. I.e., one Type1 error. AKA a 'false discovery'

Adjustment for multiple testing: Impact on sample size

Several schemes that adjust for multiple testing

E.g., Bonferroni adjustment for k tests: $\alpha' = \alpha/k$.

if α=0.05 and k=20, then $\alpha' = \alpha/k = 0.05/20 = 0.0025$.

You plan a 2-group RCT with continuous outcomes

. 80% power, α=0.05, 80% retention

. Power to detect a standardized effect size $|d| \geq 0.20$

. With no adjustment for multiple testing (α'=0.05): n=491/group

. W/ adjustment for multiple testing (α'=0.0025): n=934/group

. About a 90% increase over n=491/group

For k=5, α'=0.01: n=730/group. About a 50% increase over n=491/group

Public health contexts where multiple testing is raised

In public health, multiple testing is not at all a universal concern

It can be a concern of proposal reviewers and/or journal editors/reviewers
. Two very different audiences. More on that later

Usually raised in the context of
. RCTs
. Large-scale multiple testing situations (GWAS studies)

I have rarely seen a referee request α adjustments for, e.g.,
a regression models fit to data from an observational study

What constitutes 'multiple testing' in the context of RCTs?

Multiple outcomes

. RCT proposing to test intervention effects on multiple outcomes

RCT with >2 experimental groups and with >[#*groups*-1] comparisons

Example
RCT with two active interventions (groups A & B) and one control (C)

The plan is to perform all *k*=3 pairwise comparisons between groups.

RCTs with >2 groups and w/ exactly [#*groups*-1] comparisons planned

RCT with two active intervention (groups A & B) and one control (C)

The plan is to test A v C and B v C

A rarer and perverse perspective

Main focus: 2-group RCT with multiple primary outcomes

Example

The Community of Voices (COV) RCT. Julene Johnson, PI

Community choirs to improve the health of diverse older adults

Hypothetically, singing in a choir is a multi-modal intervention

. Cognitive: ↑ memory, executive function

. Physical: ↑ lower body & core strength, balance, lung/breath control

. Social/emotional: ↑ joy & interest in life, ↓ loneliness & depression

Case against multiple testing adjustment in RCTs: Overview

Context

- Limited set of inter-related, yet clinically distinct outcomes
- Clear hypotheses stated for each
- Transparent and honest reporting of results including
  Point estimates, CIs, and exact p-values

Adjustments for multiple testing…

- Stem from an *inductive behavior* perspective better suited to statistical process control than describing evidence from a RCTs
- Presume a universal 'null' hypothesis

Case against multiple testing adjustment in RCTs: Inductive Behavior: The Neyman-Pearson perspective

Neyman-Pearson perspective is focused on *inductive behavior*

. Decision making in repeated testing situations & taking action

This perspective is the darling of statistical process control

. Example: QC via repeated testing of widgets from production line

. Decision: Whether halt production and take remedial action

α is the long-run probability of making a Type I error

. I.e., halting the production line when there is no production problem

Neyman-Pearson focus: decisions/acting upon the evidence

. Inductive behavior: choose either H_0 or H_A

. Not about inference or generalizing from the experiment to the world.

Case against multiple testing adjustment in RCTs: Inductive Behavior: The Neyman-Pearson perspective

In the Neyman-Pearson perspective, exact p-values are not of interest

. Of interest: Whether the *p*-value is above or below α

. Given α=0.05, *p*=0.04 is regarded no differently than *p*=0.0001

Foundational tenet of Neyman-Person perspective

. Experiments will be repeated numerous times,
each time drawing from the identical population

. Across replications α reflects the expected number of Type I errors

Many have questioned its relevance to behavioral research,
where replication is very rare

Fischer regarded Neyman-Pearson a non-scientific, i.e.,
focused on decision making and not scientific inference

Case against multiple testing adjustment in RCTs: The universal null hypothesis

The null hypothesis holds for all outcomes, simultaneously

Outcomes are distinct & relevant differing facets of intervention content

We cannot prespecify which outcome or outcomes will most influence subsequent intervention-related policy decisions.

The universal null is not a good choice, usually not of interest

If you adjust for multiple comparisons, then

The decision space of the experiment should match the

decision space of anyone who might apply its results

Scientists usually can't know the decision spaces of policy makers

Case against multiple testing adjustment in RCTs: The universal null hypothesis

Many have argued that the universal null is not a good choice for RCTs †

> "The fact that a probability can be calculated for the simultaneous correctness of a large number of statements does not usually make that probability relevant for the measurement of the uncertainty of one of the statements" (D.R. Cox, 1965; p. 224)

Instead, conduct marginal (separate) tests and make marginal inferences.

I.e., a specify a test-wise error rate (e.g., $p<0.05$)

Consonant w/ Fisher: statistical tests are a tool for *inductive inference*

Marginal *p*-values represent 'strength of evidence' against individual null hypotheses

† Cook & Farewell 1996; D.R. Cox 1965; Perneger 1998; Rothman 1990

Where adjustments for multiple testing seem appropriate

Large-scale multiple testing

- . 'Mechanical' searches with no opportunity for rapid replication
- . Not (very) theory-informed or hypothesis driven
- . Null-ish relationships may be highly prevalent
- . Many tests expected to be reasonably independent of each other
- . Expect 'large' number of 'false discoveries'

<u>Examples</u>

- . GWAS looking for associations between SNPs and breast cancer
- . Swedish study looking at associations between living within 300 feet of a high-power line and 800 ailments over 25 years
- . Bible code phenomenon: groupings of words predict future events

Strategies: Peer reviewed journal articles

I have collaborated on 20 large-scale RCTs of behavioral/clinical interventions conducted in community or clinical settings

2 of 20 sets of critiques initially insisted on adjustment for multiple testing

#1. RCT of the COV intervention

. Request from a reviewer and the editor

. *The Journals of Gerontology, Series B: Psychological Sciences*

#2. RCT of a multi-modal lifestyle intervention to reduce risk of DM

. Request from a reviewer

. *AJPH*

In both cases, I wrote a response to reviewers explaining our outright rejection of adjustments for multiple testing in the clinical trial. In both cases, I prevailed

Strategies: Peer reviewed journal articles

The response included a summary of arguments presented here, plus

The quote of D.R. Cox (1965) provided previously

As well as the following quotes

Quote 1

"A motivation for much of the discussion has been the view that a clinical trial is not primarily a decision-making process, but rather a scientific experiment. Although an experiment will influence subsequent behaviour, the dependence of this behaviour on the evidential results of the trial may not be easily prespecified. The strength of evidence regarding various scientific questions may have major effect. Thus, the utilization of marginal test results and marginal *p*-values as inputs for a process of inductive inference is more consistent with this approach. Furthermore, the process of inductive behavior implied by the Neyman-Pearson framework is somewhat unrealistic given the wide variety of other factors that will influence clinical decision-making regarding an experimental treatment. The simple fact that treatment recommendations are often based on both clinical and statistical significance indicates that statistical evidence is not sufficient in itself to influence behaviour." Cook & Farewell (1996; p. 106)

Strategies: Peer reviewed journal articles

Quote #2

"The central idea behind this assertion is that, for well-defined null and alternative hypotheses, we have the capacity to interpret test results marginally and to draw inferences accordingly. The concern is that testing strategies are frequently adopted to control the overall error rate at the expense of obscuring and losing focus of the clinical questions of main interest. To reiterate Cox's (1965) comment, the simultaneous correctness of many statements does not necessarily need to be considered when focusing on a particular response." Cook & Farewell (1996; p. 108).

Cook RJ and Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. *Journal of the Royal Statistical Society, Series A*, 1996;159:93-110.

See also

Cox DR. A remark on multiple comparison methods. *Technometrics*, 1965;7:223-224.

Perneger TV. What's wrong with Bonferroni adjustments? *British Medical Journal*, 1998;316:1236-1238.

Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology*, 1990;1:43-46

Strategies: Peer reviewed journal articles

2 of 20 sets of critiques initially insisted on adjustment for multiple testing

Editorial Responses

#1. RCT of the COV intervention (reviewer and editor)

. Reviewer was completely satisfied

. Editor *said* he was not convinced, but agreed to let us publish without adjusting for multiple testing

. Editor asked that we distill our argument into a few sentences so that readers could judge for themselves.

"We did not make alpha adjustments for testing the set of clinically distinct outcomes pertinent to the hypothesized mechanisms of this multi-modal experimental intervention. Such adjustments presume a universal null hypothesis that holds for all outcomes simultaneously but, because we cannot prespecify which outcome or outcomes may most influence subsequent [<intervention name>]-related policy-decisions, the universal null is not of primary interest [REFS on previous page]. Instead, we specified a test-wise error rate to allow marginal inferences. This, in combination with the reported effect size estimates, should allow readers to draw conclusions about the impacts of the [<intervention name>] intervention on the modeled outcomes."

Strategies: Peer reviewed journal articles

2 of 20 sets of critiques initially insisted on adjustment for multiple testing

Editorial Responses

#2. RCT of a lifestyle intervention to reduce risk of DM (reviewer)

The AJPH editor accepted our argument and
accepted the paper for publication without second review

Strategies for Proposals

With a multi-modal intervention, proposal reviewers are usually OK with selection of a single primary outcome per 'mode'

A conservative approach selects a single outcome for the entire study. However, may be seen as disingenuous for multi-modal interventions

For designs w/ >2 intervention groups, some 'safe-harbor' approaches…

. If # comparisons >#*groups*-1, then adjust α for multiple comparisons

. If # comparisons =#*groups*-1, then probably OK, w/out α adjustment
Consider specifying planned comparisons as orthogonal contrasts

Ethical considerations

My perspective

RCTs of theory-based interventions with hypothesized outcomes do not require adjustment for multiple outcome testing

If you agree with this perspective,

then adjustment for multiple testing is unethical. I.e.,

. All else being equal, adjustment for multiple comparisons requires recruitment of a larger sample of participants and puts a larger number of participants at risk than is necessary

Additional thoughts

Data analysis in public health research has become highly ritualized [†]

. Select a null hypothesis of no difference, zero correlation, etc.

. Specify α=0.05. Test. If significant, accept your hypothesis

. Use this and only this procedure

Adjustment for multiple testing is not a universal part of this ritual, but

Those who insist on its universal application in RCTs are taking a highly prescriptive perspective, implying that…

. Consumers of the research can't be trusted to exercise critical judgement when integrating the results of an RCT,

and/or

. Scientists are not trustworthy, so prescriptive rituals are required

[†] Gigerenzer, Krauss, & Vitouch (2004). The Null Ritual: What You Always Wanted to Know About Significance Testing, but were Afraid to Ask. In D. Kaplan (Ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences.

Additional thoughts

Be transparent in reporting and let the reader decide

Point estimates, SEs and/or CIs, exact *p*-values

If you don't adjust for multiple comparisons and
readers want to apply adjustments,
then it is easy for them to make the adjustments

Part II

"The fact that a probability can be calculated for the simultaneous correctness of a large number of statements does not usually make that probability relevant for the measurement of the uncertainty of one of the statements" (p. 224)

Cox DR (1965). A remark on multiple comparison methods. *Technometrics*, 7, 223-224.

END