Suggested Language to Respond to Critiques Asking for Alpha Level Adjustments for Multiple Testing

Steve Gregorich, Ph.D.

Introduction

So-called 'multiple testing' can come in the form of
i. accommodating multiple outcome variables, with one regression model per outcome
ii. accommodating a categorical *X* variable with >2 categories
iii. both

Regarding *ii*, Some reviewers believe that categorical *X* variables with more than 2 categories (*c*>2) allow for *c*-1 (i.e., the *X* variable degrees of freedom) *X*-variable contrasts before adjustment to the alpha level is required. Other reviewers believe that the alpha level should be adjusted whenever the number of comparisons for any *X* variable exceeds unity. That latter position is particularly perverse, in my opinion.

Below, I provide some example responses to reviewers and corresponding manuscript text that I and my colleagues have used (successfully) to counter reviewer comments that alpha adjustments for multiple comparisons are required.

This document is organized in three parts:

Part A. Response to reviewer critique in the context an RCT
Part B. A shortened version of the essay that could be included in the main manuscript
Part C. Response to reviewer critique in the context of an observational study.

A. Response to reviewer critique in the context an RCT

The following essay argues that correcting alpha levels for multiple testing is not necessary in hypothesis-driven scientific research. I wrote the essay in response to an anonymous reviewer's critique of a manuscript that described results of an RCT with multiple primary outcome variables. We included the entire essay in the response to reviewers. The editor accepted the paper without sending out for another review.

A.1. Anonymous reviewer critique in the context of an RCT with multiple primary outcome variables

"Measuring changes in so many variables should be adjusted using the Bonferroni or similar method to avoid type I errors (false positive findings). An alpha of 0.05 across 20 variables is likely to show at least one having statistical significance."

A.2. Our response

Our goal is to communicate information about the data-based evidence from the trial. This randomized trial targeted a limited set of inter-related, yet clinically distinct, outcomes, with clear a priori hypotheses about each. In this context, we reject Bonferroni and related corrections for multiple testing for two primary (as well as many other) reasons: such corrections (i) presume the 'universal' null hypothesis and (ii) derive from an *inductive behavior* perspective of scientific inquiry that is better suited to decision-making in process control than dissemination of evidence from clinical trials.

Bonferroni and related corrections presume a 'universal' or 'general' null hypothesis, i.e., the experiment-wise error rate is concerned with a null hypothesis that holds for all outcomes, simultaneously. In the context of this randomized trial, the universal null hypothesis is not a good choice, and more importantly it is not of interest (Cook & Farewell, 1996; D.R. Cox, 1965; Perneger, 1998; Rothman, 1990). "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). The targeted trial outcomes are clinically distinct and are relevant to different aspects of 'messaging' within the experimental intervention. Thus, we sought to test and describe outcome-specific results. Our approach was to conduct *marginal* (separate) tests of each outcome and to make marginal inferences (i.e., on an outcome-by-outcome basis). Under these circumstances it is reasonable to specify a test-wise error rate, as we have ($p < .05$; Cook & Farwell, 1996; Perneger, 1998; Rothman, 1990). This is consonant with R.A. Fisher's perspective that statistical tests are a tool for *inductive inference*; here, the marginal *p*-values represent 'strength of evidence' against individual null hypotheses (Cook & Farewell, 1996; Fisher, 1973; Lehman, 1993; Perneger, 1998). This evidence can inform subsequent policy-related decisions that also incorporate much broader contextual factors (e.g., population needs, organizational capacity).

In contrast to Fisherian inductive inference, the Neyman-Pearson perspective has been described as one focused on *inductive behavior* (Cook & Farewell, 1996; Lehman, 1993; Neyman, 1961; Perneger, 1998). The inductive behavior orientation, at least originally, was more focused on decision making in repeated testing situations and acting upon those decisions (e.g., accepting or rejecting successive lots of widgets from a production line). The Neyman-Pearson perspective was decidedly a decision-focused one, emphasizing accepting or rejecting the null hypothesis (later on, they changed 'accepting' to 'failing to reject') whereas the Fisherian perspective focuses more on assessing evidence (Cook & Farewell, 1996; Lehman, 1993; Perneger, 1998).

Cook and Farewell (1996) well summarize the main points.

> "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." (p. 106)
>
> 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." (p. 108).

We do value the Neyman-Pearson perspective; with its emphasis on type I and type II errors it is key to study planning, i.e., statistical power estimation. We also believe that there are applications where corrections for multiple testing are appropriate, e.g., atheoretical, mechanical searches for relationships between health outcomes and 100s of single-nucleotide polymorphisms. However, given the context of our randomized trial and our current primary goal of communicating the strength of data-based evidence from it, we have deliberately adopted a perspective more closely aligned with Fisher.

In conclusion, we agree with Perneger (1998): "…simply describing what was done and why, and discussing the possible interpretations of each result, should enable the reader to reach a reasonable conclusion without the help of Bonferroni adjustments" (p. 1237).

References to Part A

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.

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

Fisher RA. *Statistical Methods and Scientific Research*. New York: Hafner, 1973.

Lehman E. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, 1993;88:1242-1249.

Neyman J. Silver Jubilee of My Dispute With Fisher. *Journal of the Operations Research Society of Japan*, 1961;3:145-154.

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.

Part B. A shortened version of the essay that could be included in the main manuscript

For another paper describing a different RCT, we received critiques from a review and the editor that adjustments for multiple testing must be applied. We submitted a response to that critique similar to the essay above. The editor indicated that he did not agree with our argument but agreed to publish if we added a corresponding statement in the manuscript. We added the text in section B to the manuscript.

B.1. Text added to the manuscript at the end of the Power Analysis subsection of the Methods
We did not make alpha adjustments for testing the set of clinically distinct outcomes pertinent to the hypothesized mechanisms of the experimental intervention. Such adjustments presume a universal null hypothesis that holds for all outcomes simultaneously, but "[t]he 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). Because we cannot prespecify which outcome or outcomes may most influence subsequent <*intervention name*>-related policy-decisions, we specified a test-wise error rate to allow marginal inferences (Cook & Farwell, 1996; Perneger, 1998; Rothman, 1990). 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.

References to Part B

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.

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.

Part C. Response to reviewer critique in the context of an observational study.

The following is almost identical to what is presented in Part A, with some edits to make it appropriate for an observational study. Again, the editor accepted the paper without a second round of peer review.

Part C.1. Our response

Our goal is to communicate information about the data-based evidence from this investigation. This longitudinal observation study targeted two inter-related, yet clinically distinct, outcomes. In this context, we reject Bonferroni and related corrections for multiple testing for two primary (as well as many other) reasons: such corrections (i) presume the 'universal' null hypothesis and (ii) derive from an *inductive behavior* perspective of scientific inquiry that is better suited to decision-making in process control than dissemination of evidence from public health research studies.

Bonferroni and related corrections presume a 'universal' or 'general' null hypothesis, i.e., the experiment-wise error rate is concerned with a null hypothesis that holds for all outcomes, simultaneously. In the context of this longitudinal observation study, the universal null hypothesis is not a good choice, and more importantly it is not of interest (Cook & Farewell, 1996; D.R. Cox, 1965; Perneger, 1998; Rothman, 1990). "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). Again, the targeted outcomes are clinically distinct; thus, we sought to test and describe outcome-specific results. Our approach was to conduct *marginal* (separate) tests of each outcome and to make marginal inferences (i.e., on an outcome-by-outcome basis). Under these circumstances it is reasonable to specify a test-wise error rate, as we have (p <.05; Cook & Farwell, 1996; Perneger, 1998; Rothman, 1990). This is consonant with R.A. Fisher's perspective that statistical tests are a tool for *inductive inference*; here, the marginal *p*-values represent 'strength of evidence' against individual null hypotheses (Cook & Farewell, 1996; Fisher, 1973; Lehman, 1993; Perneger, 1998). This evidence can inform subsequent policy-related decisions that also incorporate much broader contextual factors (e.g., population needs, organizational capacity).

In contrast to Fisherian inductive inference, the Neyman-Pearson perspective has been described as one focused on *inductive behavior* (Cook & Farewell, 1996; Lehman, 1993; Neyman, 1961; Perneger, 1998). The inductive behavior orientation, at least originally, was more focused on decision making in repeated testing situations and acting upon those decisions (e.g., accepting or rejecting successive lots of widgets from a production line). The Neyman-Pearson perspective was decidedly a decision-focused one, emphasizing accepting or rejecting the null hypothesis (later on, they changed 'accepting' to 'failing to reject') whereas the Fisherian perspective focuses more on assessing evidence (Cook & Farewell, 1996; Lehman, 1993; Perneger, 1998).

In the context of clinical trials, Cook and Farewell (1996) well summarize the main points.

> "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." (p. 106)

> 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." (p. 108).

We do value the Neyman-Pearson perspective; with its emphasis on type I and type II errors it is key to study planning, i.e., statistical power estimation. We also believe that there are applications where corrections for multiple testing are appropriate, e.g., atheoretical, mechanical searches for relationships between health outcomes and 100s of single-nucleotide polymorphisms. However, given the context of our investigation and our current primary goal of communicating the strength of data-based evidence from it, we have deliberately adopted a perspective more closely aligned with Fisher.

In conclusion, we agree with Perneger (1998): "…simply describing what was done and why, and discussing the possible interpretations of each result, should enable the reader to reach a reasonable conclusion without the help of Bonferroni adjustments" (p. 1237).

References to Part C

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.

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

Fisher RA. *Statistical Methods and Scientific Research*. New York: Hafner, 1973.

Lehman E. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association*, 1993;88:1242-1249.

Neyman J. Silver Jubilee of My Dispute With Fisher. *Journal of the Operations Research Society of Japan*, 1961;3:145-154.

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.