Statistical Adjustment for Multiplicity Virtual Workshop
October 26-27, 2022

Keynote Presentation and Speaker Sessions
Moderated Panel and Question-and-Answer Discussions

Overview

Leading scientists from academia, industry, and government in the fields of epidemiology and statistics came together for the first National Cancer Institute (NCI) workshop to consider adjustments for multiplicity that directly impact the statistical significance of research results and drive change in how studies involving multiple comparisons are conceived, reported, and ultimately, understood as a body of work. The goals of this meeting were to:

- Review evidence from the field to assess how multiplicity adjustments are being implemented (or not)
- Foster idea exchange from traditional, new, and emerging viewpoints regarding multiple testing procedures (MTP) and their application to improve concepts and methods while considering a more unified approach
- Query participant trialists on whether and how to adjust for multiplicity within primary, secondary, and exploratory research

Day 1—Wednesday, October 26, 2022

Welcome and Introductory Remarks

Philip E. Castle, PhD, MPH
Director, Division of Cancer Prevention (DCP), NCI

Via a recorded message, Dr. Castle reminded attendees about the limitations of population studies in cancer research, arising from the relatively rare incidence of cancer in healthy populations and the finite resources to perform research. Since data from humans are often noisy, Dr. Castle emphasized that statistical thinking is needed to uncover meaningful signals.

Dr. Castle underscored the need for enhanced rigor and reproducibility. However, the requirement to minimize sampling error will necessitate better precision in measurement, more efficient statistical methods, or larger sample sizes, requiring a balance to be struck. Although larger samples can reduce the need for extra confirmatory research, broadening a study's scope will require more participants and resources, which poses over-recruitment and treatment risks.

Dr. Castle explained how researchers can apply "intentional" study design to control for multiplicity, particularly when limited population sizes are involved. He recommended that researchers home in on the following fundamental questions when designing research:

- What scientific questions are you attempting to answer?
- What factors will be controlled for?
- What statistical methods will be applied?
- How will rejecting a true null hypothesis, and failing to reject a false one, be handled?

Dr. Castle welcomed participants to engage in further discussion and collaboration on these important statistical issues.

Keynote: A Broad View of Multiplicity Adjustment

Selective Inference Issues in Medical Research: Science and Politics

Yoav Benjamini, PhD
Statistics and operations researcher, School of Mathematical Sciences, The Sagol School of Neuroscience, The Safra Center for Bioinformatics, Tel Aviv University

As a result of his examination of published cancer research and evolving professional guidelines, Dr. Benjamini catalogued the current avenues of thought regarding "selective inference," or the principle that when further hypotheses or findings are selected after reviewing the data, statistical rigor deteriorates, and hence, replicability becomes compromised.

Dr. Benjamini allowed that selective inference is unavoidable. Yet, this existence is not something he is against. He outlined a procedure designed to control the false coverage rate, which, if followed, could mitigate the deleterious effects of selective inference. Also, he proposed that researchers remain mindful of multiplicity and, in doing so, consider how revised experiments and methods can reduce any negative influences of selective inference. In this manner, not only primary outcomes but also secondary and exploratory endpoints can provide meaningful information to advance an area of study.

Dr. Benjamini framed selective inference as the driver for the meeting's three sessions, intended to discuss the philosophy of, and find solutions in, the application of statistical adjustments. In jump-starting the meeting's pursuits, he relied on published research (and study proposals he has received) to illustrate some current quandaries, such as:

- Early findings from one multifactor test were statistically significant, but later studies on the same topic indicated a clear drift into statistical insignificance.
- Once published studies fail on statistical significance for the primary endpoint, the secondary and exploratory endpoints are not considered or reported.
- Statistical significance was not applied to secondary or exploratory endpoints.

Dr. Benjamini emphasized the importance of adjusting sample size and MTP power for multiplicity, regardless of the number of tests in a study. This adjustment must first be specified and then calculated. For example, this could be the expected proportion of true signals that are discovered (average power) or the probability that a given proportion of true effects are discovered (TPX power). He remarked that there is an insufficient body of published work on multiple test power and encouraged work in this area.

Dr. Benjamini also discussed a 2016 editorial in the *Journal of the American Statistical Association* about p-values. Although finding the p-value has been recognized by professional associations and publications as useful in representing statistical significance, this essay admonished what it called "widespread misuse of the p-value" and insinuated that inferential procedures based on it were to blame. Following this editorial, several medical journals changed their guidelines for statistical reporting. In 2020, an American Statistics Association task force, of which Dr. Benjamini was a member, recommended that the p-value be applied properly rather than abandoned.

As Dr. Benjamini shared, not only does the p-value remain the first defense against false positives, but also this testing method for statistical significance is widely used across all the sciences. He made

recommendations for statisticians and practitioners to further consider and apply statistical adjustments, such as:

- Ordering selection rules effectively
- Applying hierarchy and weighting to primary, secondary, and exploratory endpoints to include those outcomes, and possibly using a weighted false discovery rate (FDR) procedure to threshold all hypothesis tests from all outcomes groups of a trial simultaneously
- Implementing the method of false coverage rates for the confidence interval (CI), which employs marginal intervals away from zero and adjusted intervals toward zero

Session 1: When to Adjust for Multiplicity

Moderated by Victor Kipnis, PhD
Chief, Biometry Research Group, DCP, NCI

Adjusting for Multiplicity in Clinical Trials and Observational Studies Is Critical and Absolutely Must Be Done* *Only if You Plan on Turning Off Your Brain and Blindly Following a Dichotomized P-value

Andrew J. Vickers, PhD
Attending research methodologist, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center

Dr. Vickers encouraged researchers to look beyond the limitations of the p-value figure as a rigid line between study conclusions that are deemed to be "gospel truth" and those that are relegated to fallacy. However, he explained, while p-values are problematic, they should not be abandoned in the same way a bicyclist rides, knowing it rains sometimes or an expensive wheel rim might be damaged.

Dr. Vickers began by saying he was not speaking about clinical trials, for which there is a regulatory mandate to adjust primary outcomes for multiplicity, nor was he speaking about basic science (e.g., animal studies or MTP in high-dimensional omics settings), for which there is an obvious need for multiplicity adjustment. Rather his remarks were primarily focused on observational epidemiology and exploratory outcomes in clinical trials.

The first step is to consider whether adjusting for multiplicity is a necessary prerequisite to ending up with useful findings. The accompanying question is whether test results should be dismissed out-of-hand for p-values that fall under the accepted threshold.

Dr. Vickers asked rhetorically, from a historical perspective, whether some key studies should have been corrected for multiplicity, such as the 1950 study on cancer and smoking. The landmark work by Doll and Hill did not adjust for multiplicity at all and could be heavily criticized for not doing so. He pointed to literature that showed no lack of later studies, which confirmed the initial findings, continuing to build evidence on the health and mortality dangers of smoking in the decades that followed. If the original work had not been published, none of the confirmatory work would have followed. So, if the p-value is the ultimate answer, then why, Dr. Vickers asked, would researchers ever perform more than one trial? He also wondered about the efficiency of doing so, when so many individual papers spring from one study involving multiple tests.

To perhaps illustrate a more pragmatic approach to data gathering, Dr. Vickers paraphrased philosopher David Hume: "I need more evidence there's a unicorn in my garden than there's a horse." Further, when endpoints are highly correlated, he explained, the rest of the data should not be thrown out only because the primary endpoint fails on p-value. He remarked that the Bonferroni procedure is too conservative because the individual hypothesis tests are correlated. A solution he proposed, but has not seen yet, would be to create a covariance matrix, but what factors would be used, he asked, and isn't this meta-analysis anyway?

Dr. Vickers concluded that multiple testing adjustments should be rarely applied in the setting of observational epidemiology and by extension, the analysis of exploratory outcomes in a clinical trial—and, possibly, not even applied in the analysis of secondary endpoints in clinical trials. In addition to paying attention to the number of hypotheses tested, he pointed to the following criteria for a more nuanced picture of the efficacy of a trial or an intervention:

- Methodological quality
- Supportive evidence in the literature from similar studies
- Biological plausibility from basic science
- Consistency of findings within the study
- Strength of evidence against the null

In response to Dr. Vickers' assignment of a good portion of the Bonferroni method's excess conservatism to its independence assumption, Grant Izmirlian, PhD, presented a diagram with a plot for discussion. Derived from the multivariate normal distribution, this diagram plotted the Bonferroni correction versus correlation for two hypothesis tests; the results indicated that excess conservatism for the independence variable is only noticeable for correlations in excess of 0.75.

Pointing to his example of a trial with many interconnected endpoints (with some highly correlated), Dr. Vickers said that the concern is broader than Bonferroni being open to excess conservatism. It's that we don't know what the adjustment should be or whether uncorrelated results indicate more about the questionnaire than about treatment efficacy. As statisticians, he continued, we apply precise corrections, as we typically do; however, we ignore the issue of correlation.

Multiplicity of Inferences in Medical Research: Accounting and Reporting

Constantine Gatsonis, PhD
Henry Ledyard Goddard University professor of statistical sciences, Department of Biostatistics, director of the Center for Statistical Sciences, Brown University

Dr. Gatsonis framed his presentation by asking whether recent scrutiny of multiplicity from the lens of reproducibility has made a difference in how research results are interpreted and, likewise, reported in the literature. Most importantly, this recent scrutiny has highlighted the need for conservative reporting of study results. Multiplicity has long received attention in the clinical trials literature and has led to an extensive and nuanced body of frequentist theoretical research and computation. However, it is considerably less developed within the Bayesian paradigm. Well-known statistician James O. Berger has published some very thoughtful work along this line.

Dr. Gatsonis advised that addressing multiplicity considerations must start with the study design. In framing the issue, he proposed that (1) the decision on whether to adjust for multiplicity should be made, based on consideration of study type, and (2) the multiplicity control approach should be pre-specified.

For example, there are several possibilities for defining families of tests in a clinical study, such as:

- Controlling the study-wide error (e.g., primary, secondary, and exploratory endpoints) treated together, adjusting for the multiplicity of all reported analyses or comparisons in one fell swoop.
- Conducting group comparisons, blocked within levels of a chosen variable, controlling only the group-wide error. For example, a variable is post-hoc chosen to define groupings or families of tests, and multiplicity is adjusted for separately within levels of this variable.

OR

- Defining subgroups of interest (e.g., primary, secondary, and exploratory endpoints) and controlling the error rates separately within each of these groups or families.

Dr. Gatsonis pointed out that the first and third options are defensible, but the second one leaves open the potential for manipulation. Considering the overall strategy is extremely important, he stressed. In a confirmatory study, the control of Type-I (family-wise) error may be needed, whereas in an exploratory study, control of FDR may be more appropriate.

Next, per the session title, Dr. Gatsonis initiated a discussion of which results should be reported in a published manuscript. Should only controlled comparisons be highlighted in the abstract, along with discussion of all comparisons of interest? He decried the reporting of nominally significant findings and noted that most journals are moving away from allowing it.

Accounting for multiplicity of tests, along with CIs, is a standard aspect of regulatory decision-making, and adjustments are often prespecified in randomized clinical trial (RCT) design to prevent surprises in the data. However, in a study on a set of 500+ cardiovascular RCTs that Dr. Gatsonis cited, less than 30% controlled for multiplicity. For observational studies, he found the application of multiple testing controls rare. Therefore, he noted the difficulty in imposing a standard—no matter what it is—and asked two questions:

- Should only controlled comparisons be reported or all comparisons of interest?
- If one were to report "nominally significant" findings, is the p-value useful or misleading?

Interestingly, Dr. Gatsonis connected the use of a CI for hypotheses testing to a reasonable requirement to apply the same practice to multiple tests. And to help resolve the significant/not significant dichotomy, he suggested that a CI has a role when researchers summarize evidence.

As an example of a journal board that has established author guidelines. Dr. Gatsonis pointed to the *New England Journal of Medicine.* Its guidelines state the following:

- Prespecified multiplicity methods are essential.
- When no multiplicity methods exist, reporting on secondary/exploratory endpoints should be limited to point estimates of effects with 95% CI.
- P-values for multiplicity should be labeled as such.

- P-values should not be reported after the first insignificant finding.
- For exploratory research, investigators may use FDR.

Although it is clear that accounting for multiplicity of inferences is occurring and conscientiousness has been raised, Dr. Gatsonis warned that this momentum is not guaranteed. In particular, he advised that methods and consensus on approaches for Bayesian paradigm be further developed.

During the question-and-answer period, Dr. Kipnis inquired about Dr. Gatsonis's recommendation to apply family-wide error rate (the probability) when testing primary/secondary endpoints while employing the less-stringent FDR (mean for false discovery) for exploratory results. Dr. Gatsonis responded he was open to a less-stringent alpha level to define how strong the evidence must be to contradict the null hypothesis.

Multiple Comparison Controversies Are about Context and Costs
(The Need for Cognitive Science and Causality in Statistics Teaching and Practice)

Sander Greenland, DrPH, MA, MS
Emeritus professor of epidemiology and statistics, Department of Epidemiology, School of Public Health, University of California, Los Angeles

Dr. Greenland described prevailing biases in the teaching and practice of statistics and the need to understand the impact of the differing stakeholder values on how research is conducted and perceived, particularly in the observational study of multiple factors at one time. He then made recommendations for reform that will better help researchers pool findings for meta-analysis and, thereby, build bodies of evidence.

Dr. Greenland proposed that null hypothesis significance testing can lead to censoring of information and can prevent the development of unbiased public data repositories. He also noted:

- An underlying value bias exists toward accepting the null hypothesis (proposing no relationship between an agent and the effect being studied).
- Ambiguous results are typically reported as negative findings.

Because results are seen as an either–or dichotomy, a study can quickly be deemed unpublishable. Further, Dr. Greenland stated many studies are being performed on an agent because a known effect exists, so why start with an assumption that one does not?

The prevailing belief that false positive costs are higher than false negative costs fails to consider that companies and patients/consumers come to research desiring opposing outcomes; for example, although a drug company monitoring adverse effects wants no problems, a patient wants problems to be uncovered. Another example he gave was about a chemical company desiring a harmless (noncarcinogenic) finding for its new product, whereas consumers want to understand the cancer risk. Whether the risk is small enough is a value judgment.

Dr. Greenland recognized that multiple comparison corrections serve decision-oriented goals (e.g., selection for further study). Therefore, the most-defensible research goals are to (1) describe how data are generated and (2) summarize the data and the information they relay. Therefore, a reporting

practice that brings all assumptions forward in unconditional descriptive language should be encouraged.

For the cancer research community, Dr. Greenland made key recommendations:

- Describe the decision-based goals and benefit–loss structures to justify the test hypothesis, the use of multiple-comparison corrections, and the p-value cutoff/Bayes factor.
- Understand that the results of multiple-comparison adjustments may be justifiably rejected by stakeholders having different goals.
- Never supplant true information (in the form of data) with the results of multiple-comparison correction procedures.
- Rather than for multiple comparisons, include the p-value when a single factor is being tested in isolation.
- Develop additional work on the Bayes approach (a mathematical means of incorporating prior beliefs and evidence to produce new beliefs).

During the question-and-answer period, Dr. Kipnis broached a comment on Dr. Greenland's observation of researchers frequently using the null hypothesis as the test hypothesis (alternative), stating that the word "null" in terms of hypothesis can be misread as "null effect." Dr. Kipnis stated that the null hypothesis is simply the hypothesis the researcher is testing, leaving open the possibility for a change in viewpoint, based on what the data indicate. Dr. Kipnis added that rejection or acceptance of the null hypothesis signals the opposite decision for the alternative hypothesis.

Session 1: Question-and-Answer Panel

Andrew J. Vickers, PhD
Constantine Gatsonis, PhD
Sander Greenland, DrPH, MA, MS

Joined by:

Joseph Unger, PhD, MH
Associate professor of biostatistics, health services researcher, Cancer Prevention Program, Public Health Division, Fred Hutchinson Cancer Center

Prompted by questions from workshop participants, the panelists discussed the concepts that drive when and how multiple comparison corrections should be applied.

If you have highly correlated tests, the Bonferroni method is too conservative; so what should the adjustment be with this method?

Based on his team's experience creating stopping rules, in which p-value and alpha are observed, Dr. Vickers understands the statistician's practice of being precise. Perhaps, it's not that Bonferroni is too conservative; rather, the adjustment for correlated tests are unknown.

How can a paper be salvaged for publishing in journals, all with different multiplicity requirements, when the approved trial design has not yet dealt with this issue?

Dr. Gatsonis made recommendations for this difficult situation. Acknowledging that the journal will likely publish some form of the paper—perhaps only the primary endpoints—he advised submitting a data description and writing from the patient perspective. He added that corrections are not always required, such as for safety endpoints. In understanding the rationale for publishing all p-values, Dr. Gatsonis asked researchers to see where these p-values may end up—in ads for a product or treatment somewhere. He also explained that guidelines help corral consensus on a topic.

Dr. Vickers countered with a strong statement that journals should not dictate to researchers what methods to use; and Dr. Greenland added that by publishing requirements, journal boards seek to dictate values.

What is the No. 1 thing you would change in teaching for the next generation to move the field forward?

Dr. Greenland advocated for reform in teaching practices, starting in the basic statistics class, while understanding the distinction between both the need to smooth data for purposes of summarization and the imperative need to build loss (cost–benefit) structures for decision analysis. This lively back-and-forth discussion, and the prevailing caution to be careful with adjustments, led to the following more fundamental query:

What is a conservative approach to multiple correction procedures?

Dr. Vickers asked a more basic question: What does "conservative" mean? It could mean protecting people, or it could mean protecting the status quo. In answering this core question, he concluded it's hard to translate these ideas into a statistical principle. As a clinical trialist in a collaborative program, Dr. Unger simply asked, do we trust the research community to make the appropriate determinations, or do we implement top-down control?

What are your opinions about data sharing after the data has been adjusted? Can others use the data as they like?

In addition to relaying strong support for public data (beyond adjusted data), Dr. Greenland stated it was fine for others to use adjusted data, as long as subsequent authors accurately describe what they're doing with it. He reemphasized the need to explain what values influence any claims they reach. His rationale for this position is that so much is hidden in trials, conflicting with the crucial public interest in sharing that information. Dr. Greenland proposed flipping the hypothesis from "no harm" to "there is harm" to account for stakeholders on the opposing side.

What about exploratory testing, the need to control for it or not, and the influence of exploratory findings on generating hypotheses for future studies? Would this result in lots of false positives, which waste resources, or in false negatives, which may preclude further investigation?

Dr. Andrews commented that an author can't use hypothesis generation to claim something and not take responsibility for it. He also questioned whether exploratory research is needed to generate hypotheses while acknowledging the benefit in Phase I trials, for which effects are not yet known.

Session 2: How to Adjust for Multiplicity

Moderated by: Grant Izmirlian, PhD
Mathematical statistician in the Biometry Research Group, DCP, NCI

A Quick Overview & Comparison of Methods to Adjust Tests on a Modest Number of Outcomes

Luke Miratrix, PhD
Associate professor of education, affiliate faculty in the Department of Statistics, Harvard University

Dr. Miratrix ran through the underlying definition and methods currently in practice for multiplicity corrections. He began by introducing the Bonferroni method, which relies on an "alpha" budget of 0.05 that is allocated to individual tests by either equal distribution or hierarchal percentage. Dr. Miratrix reviewed the definition of randomization logic; by applying the concept to a controlled trial, he framed the understanding of correction methods and the connection to the p-value. Although the distribution of covariants (e.g., age) should be balanced across the treatment and control groups, randomizing imbalance can arise, creating results that are more likely to occur by chance. If more healthy outcomes are observed in the treatment group than the control, the agent worked—or it was the result of "good luck" in the random assignment. If the treatment failed and there are false positives, the result was due to "bad luck" in the randomization.

Dr. Miratrix expounded further, explaining how the Bonferroni method reduces power for a study with multiple highly correlated endpoints by not taking advantage of the correlation. The result is an appearance of more mistakes in the data, making it more difficult to reject the null.

As a permutation approach to the Bonferroni method, Dr. Miratrix discussed applying the Westfall-Young model to simulate the p-value (and later, the t-statistic) distribution while preserving the correlation between tests. Because power level is closely tied to correlation strength, he defined "power" at three levels for the chance of rejecting:

1. The first hypothesis (individual)
2. At least one hypothesis (1-minimal)
3. All hypotheses (complete)

This model compares the actual p-value to what is expected, assuming no treatment effect:

1. Shuffle the data (1,000 times).
2. Re-estimate the p-values.
3. Record the minimum p-value across outcomes.

Dr. Miratrix noted that the Westfall-Young method has additional complexity beyond the first test, and further inquiry is needed to determine reasonable values for the above analyses. Nonetheless, because the model records the minimum p-value, there is confidence in the corrected p-value, and the estimate has a less severe effect on power than the Bonferroni result.

By running a simulation using Power Under Multiplicity Project (PUMP) software and graphing the various correction methods, Dr. Miratrix compared the chance of rejecting the outcome for several multiplicity correction methods. With the x-axis as "individual power" and y-axis as "rho" (strength of

association), at the 0.50 rho, the Westfall-Young step-down and Westfall-Young single-step models outperformed the Holmes and Bonferroni methods.

To Adjust (or Not to Adjust) for Multiplicity of Decision Paths, Endpoints, Subgroups, and Tumor Types

Jason Hsu, PhD
Emeritus professor of mathematics, The Ohio State University

With an emphasis on harmonization between the science discovery and applied research communities, Dr. Hsu gave his perspective on applying multiplicity adjustments in the industry-sponsored RCT setting—as regulated in the United States by the Food and Drug Administration (FDA) and internationally by the European Medicines Agency and the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Dr. Hsu highlighted the different views evident in drug-approval settings:

- Primary endpoints are critical to show the efficacy necessary for FDA approval.
- Secondary outcomes serve as indications for labeling purposes.

In this practice, secondary endpoints are not of import until efficacy is proven by the primary outcome. If there is only one decision path, there is no need for multiplicity adjustments.

For studies with multiple endpoints, Dr. Hsu presented an alternative to closed testing. Termed "partition testing," this approach disjoints correlations where they overlap to obtain further insight. In particular, he cited its application to analyzing dose-response data (e.g., low, medium, and high dosage), which researchers can use to uncover the effects of each endpoint and dosage combination.

Dr. Hsu also noted that the identification of patient subgroups in cancer research has progressed in alignment with the biological mechanisms found to cause or increase the risk for cancer. Although some subgroups (lacking the risk mechanism) will not be helped by a new drug regimen, other patients must first be referred for genetic or other biological testing to determine subgroup.

Dr. Hsu recommended applying a "basket" approach to protocol design. With this design approach, therapies across different types of tumor histologies can be evaluated, and the physical scarcity of patients and tissue can be addressed. He opined that the basket approach for tumor research could harmonize Bayesian theorem with the interpretation of Frequentists, who consider frequency of repeatable experiments during information gathering.

Dr. Hsu also discussed FDA's need to balance risk and benefit for the public in making sound approval decisions; therefore, the stability of error rates matters. Not only is family-wise error rate (FWER) applied within a study, FWER is also applied via study batches to connect the error rates with incorrect approval decisions.

Empirical Bayes Methodology for Multiplicity Adjustment

Daniel Yekutieli, PhD
Professor, Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University

Rather than discuss multiplicity adjustment per se, Dr. Yekutieli described the Bayes methodology through the lens of selective inference, a concept Dr. Benjamini reviewed in the first session. Using published references ranging from 1951 to 2014, Dr. Yekutieli illustrated that the problems—or solutions like FDR—are not new. He then demonstrated, step by step, how to implement the empirical-based Bayes method, albeit using a theoretical study, involving testing hundreds of compounds for positive effects.

Dr. Yekutieli demonstrated that the Bayes method seeks to combine prior information with the likelihood of an effect in a new sample to gain a posterior, or new, probability. Among other techniques used to solve problems of selective and simultaneous inference in the Bayes method (e.g., Bonferroni CIs, FWER control, Bayesian inference, false-coverage rate-adjusted CIs), marginalized losses are aggregated to account for the cost of inaccurate predictions.

In the hypothetical example, Dr. Yekutieli applied the Bayesian approach to high-dimension objects with many outcomes, using sampling as realized within the study's data. Today, Bayesians strive to correct marginal inferences for multiplicity via prespecified parameters, and the goal of the example was to find "big" positive effects from the library's compounds.

In addition to two prior distributions that were close to normal, Dr. Yekutieli relied on the following three strategies to determine the average risk for selection bias in performing further data analyses, running the experiments until the results were significant:

1. Employ random selection, and run the experiment on all compounds.
2. Select each compound, and run the experiment.
3. Randomly select a compound, and repeat the experiment for the same compound.

Assuming the effect is positive, the strategy is repeated to find a positive or negative result for each test until significant results are achieved. A set of effect estimator rules are used, with the goal of finding the strategy that minimizes risk and understanding whether more selection equals more risk:

1. Maximum likelihood estimation
2. Posterior mean for prior distribution updated by the likelihood (Bayes rule 1; shrinkage)
3. Posterior mean for prior distribution updated by the truncated likelihood (Bayes rule 2; not applied if the risk isn't lowered)

As applied in this theoretical study, Dr. Yekutieli underscored that selection does matter. By repeating the selection strategies many times, bigger data and better numerical assessments follow.

The findings were displayed on a table through each step of the Bayes analysis. For the three selection strategies, the Bayes Rule 1 estimator was the most effective in minimizing average risk of selection bias for the first two selection strategies. However, the Bayes Rule 2 estimator was more effective at minimizing the average risk for the third selection strategy. Therefore, the use of the different estimators depends on the chosen selection strategy. In applying his example to the research setting, Dr. Yekutieli recommended running a smaller exploratory analysis to map the compounds in the library, using an empirical-based distribution rather than a theoretical one.

Session 2: Question-and-Answer Panel

Luke Miratrix, PhD
Jason Hsu, PhD
Daniel Yekutieli, PhD
Joseph Unger, PhD, MH
Vance Berger, PhD
Victor Kipnis, PhD

Prompted by questions from the other panelists and attendees, speakers from Day 2 engaged in a conceptual and technical discussion of the Bayes paradigm.

> **Prior distribution seems relevant, and the Bayesian concept of updating beliefs seems appropriate. But does a new study have merits on its own? Should it be independent?**

Because his example assumed prior experience with the compounds, Dr. Yekutieli recognized the importance of considering prior experience as part of any new study. Dr. Greenland added that Bayes and Frequentism share the idea of a discovery continuum, whereby researchers average new data with prior evidence, using multilevel models to improve inferences and reduce loss. Dr. Hsu added that although it might be difficult to translate study contexts and measures into probability, researchers ought to try.

> **There is some thought that a laissez-faire approach is best for allowing researchers to deal with multiplicity. Since experts inform NCI so the Institute can develop reasonable guidelines, this may conflict with the workshop's goals. Are we shooting this goal down?**

Dr. Hsu noted that the recent FDA efforts to draft multiplicity guidelines has been a hot topic, along with ICH actions to amend its E9 statistical principles for clinical trials. In the United States, FDA guidance recommends submitters adjust for baseline covariants. Dr. Berger stated that multiplicity can't be ignored. After reviewing the drawbacks and capabilities of the conservative, Bayes, and Frequentist approaches, Dr. Kipnis emphasized it is not possible to argue convincingly "there is an effect" or "there is no effect" until studies are repeated.

Dr. Sander noted that using the Bayes strategy is the same as performing certain meta-analysis techniques. However, the fruitful practice of combining empirical Bayes and Frequentist approaches remains largely unrecognized, he added.

Emily Van Meter Dressler, PhD, submitted the following comment from the perspective of a research-based statistician at Wake Forest University: "In order for science to grow, we must adapt and explore. I would hesitate to make recommendations, lacking general consensus. Feasibility can't be compromised, yet we must leave room for out-of-the box thinking. I find it a struggle to find a happy medium."

Day 2—Thursday, October 27, 2022

Session 3: How to Power Multiple Testing Studies

Moderated by: Vance Berger, PhD
Mathematical statistician in the Biometry Research Group, DCP, NCI

Power Multiple Testing in Omics Studies

Peng Liu, PhD
Professor of statistics, Department of Statistics, Iowa State University

With a perspective gleaned from omics research (e.g., microarrays, genomics, microbiomes), Dr. Liu described the application of multiple testing controls to high-dimensional datasets, often composed of millions of snips and thousands of samples. In RNA sequencing (RNA-seq), for example, hypotheses are developed for thousands of genes. To perform differential expression analysis, genes can also be grouped into multiple pathways and networks for further investigation, based on biological information.

Dr. Liu delivered her rationale by answering the following questions, which she supported with data examples and modeling equations:

1. Why should we control for multiple testing error rate?
2. Which error rate should be controlled for?
3. How do we derive powerful tests?

Dr. Liu identified challenges unique to her setting, namely, large dimensions of variables/features measured simultaneously (i.e., small sample size, large "p" problem).

In answering the first question, Dr. Liu concluded that controlling for error rate is necessary because this correction reduces false positives. To answer the second question, she proposed FDR as the most appropriate error rate to control.

Recognizing there has been limited discussion on the third area of inquiry, she presented the following techniques to build a powerful model:

- Finding higher degrees of freedom (independent pieces of information) to give a "higher" sample size without collecting more samples
- Improving variance/dispersion estimation—underestimation for variance can create false positives, and overestimation leads to loss of power
- Finding the proper cutoff for the threshold

According to Dr. Liu, the goal is to derive power while controlling FDR. In doing so, she employed the following software packages for sample size calculation:

- Ssize.fdr—to calculate sample size for t- and F-tests while controlling for FDR
- Ssize.rna—specifically designed for RNA-seq testing

To meet the identified challenges of large experiments with multiple testing and a substantial number of parameters, she synthesized the following takeaways:

- Ensure study designs have enough sample size to guarantee power.
- Improve the power of individual tests.
- Use better estimates for model parameters.
- Compare the average power by using the maximum average power framework.

The Power Under Multiplicity Project: An R Package for Calculating Power for Multilevel Experiments

Kristen Hunter, PhD
Statistician, Department of Statistics, Harvard University

Describing a joint project with Dr. Miratrix, Dr. Hunter detailed the capabilities of PUMP software to support practitioners as they apply adjustments for single- and multilevel experiments. Most RCTs today, she explained, don't take into consideration MTPs that, when applied optimally, can help ensure study conclusions are valid. PUMP supports studies with up to three levels each of hierarchal covariate design and randomization.

Operating under Frequentist probability and assuming a mixed effects regression model, Dr. Hunter introduced how PUMP can help each researcher uniquely preview and define success in their research contexts. Because there are multiple ways of defining power in MTPs, PUMP supports a set of comprehensive definitions:

- Individual—detect per test
- 1-minimal—probability to detect one or more of the true signals/reject one or more null hypotheses
- 2-minimal—probability to detect two or more of the true signals/reject two or more null hypotheses
- D-minimal—probability to detect d or more of the true signals/reject d-number of hypotheses
- Complete power—detect all of the true signals

PUMP supports the user in selecting one of several recognized multiplicity adjustment methods:

- Bonferroni—Alpha-level adjustment for family to control for Type-1 error
- Holm—Step down for Bonferroni
- Benjamin-Hochberg (B-H)—Step up (FDR, less conservative)
- Westfall-Young—Single step and step down (not overly conservative)

PUMP also estimates optimum sample size and minimum detectable effect size (defined as the minimum numerical difference between two groups).

The software can simulate and display potential multilevel adjustments across all tests, individual mean, and all definitions of power, plotting results on a table and accompanying graph for clear visual comparison. The power result for no adjustment is also displayed for reference, and estimates can be rerun for more precision.

Users can employ the update function to revise a power call as assumptions change, without making time-consuming efforts to run the data again. Among other parameters, such as target power and tolerance, PUMP accepts parameters for intercepts that are mixed or random as well as treatment

effects that are constant, fixed, or random. PUMP helps automate the following steps, based on parameters the user enters:

1. Calculate test statistics.
2. Calculate p-values.
3. Determine power, using distribution of p-values.

If p-value is less than the alpha, the result is power.

Researchers can also assess the sensitivity of power. As an advanced feature, users can enter the necessary variables and estimate the correlation between test statistics, based on the correlation between outcomes.

Sample Size Calculations for FDR Control

Sin-Ho Jung, PhD
Professor of biostatistics and bioinformatics, Department of Biostatistics and Bioinformation, Department of Basic Science, Duke University

Dr. Jung itemized the particulars of applying sample size calculations to high-dimension data while also controlling for FDR, using a microarray testing example to locate the over- or under-expression of genes for two phenotype patient groups. Although this method may apply to any type of high-dimension data, Dr. Jung cited a study of approximately 6,800 genes being tested simultaneously for 11 leukemia patients.

After presenting a data design map, he explained the goal was to determine how many microarrays (subjects) were needed at a prespecified FDR level to detect a certain number of differently expressed genes. He noted that although FDR is more powerful and requires less computations than FWER (the other option for regulating the Type-1 error rate), FDR is harder to control accurately. The discovery procedure for this example involved 6,810 t-tests, so the error rate would be enlarged without a multiple testing adjustment. The following two hypotheses were developed for each gene:

- The gene was equally expressed.
- The gene was differentially expressed.

Input parameters to determine the proper number of microarrays included:

- Number of genes and those expressed differently
- Constant effect size of differentially expressed genes, standardized
- Allocation proportions between each group
- Desired FDR level
- Number of desired discoveries ($<m_1$)

Dr. Jung ran 5,000 simulations on the large dataset to calculate the number of subjects needed under each design setting and count the number of true rejections. Then, distributions were plotted on histograms for four simulation settings. (The estimation is good if the number of true rejections is approximately equal to the number of true discoveries.) A one-sided test was chosen to reject one side of the curve, but extension to a two-sided test would simply end up requiring more subjects. (Dr. Jung

noted an unequal effect size would require an additional numerical method to solve the equation.) He concluded that between 68 and 73 patients were needed.

Session 3: Question-and-Answer Panel

All speakers and attendees invited to join.
Peng Liu, PhD
Kristen Hunter, PhD
Sin-Ho Jung, PhD
Joseph Unger, PhD, MH
Kevin Dodd, PhD
Grant Izmirlian, PhD
Victor Kipnis, PhD

For this session, speakers and attendees gathered to dive into some technical aspects of powering studies as part of making multiplicity adjustments.

> **Do you find Storey's model for p-naught (probability) noisy?**
>
> **Where does the utility of average power break down in terms of multiplicity? When is the expected value not a good measure of what we might experience for a one-point sample in our experiment?**
>
> **In our experiments, we have 10,000 genes and 50,000 biomarkers—the sample size is often set already. We are using an application, so the sample size calculation increases the experiment's cost. How do we account for the balance between power and effect size?**

Encouraging others to define the term in their contexts, Dr. Jung defined "power" as applied in his setting: Power is the number of true discoveries wanted, divided by the number of genes expected to be differentially expressed. He also indicated that variable dependency or independency matters to a model. Because FDR assumes independence or light dependence, there is no FDR that gets it exactly right nor are there any permutation methods yet. Therefore, he encouraged statistically minded researchers to consider creating an insight-based prediction from the outset, based on their own experiences. As an alternative, Dr. Jung recommended FWER, where applicable.

Plenary Panel Discussion

Moderated by: Victor Kipnis, PhD
Chief, Biometry Research Group, DCP, NCI

All speakers and attendees invited to join.
Luke Miratrix, PhD
Joseph Unger, PhD, MS
Grant Izmirlian, PhD
Kevin Dodd, PhD

Joined by:
Dan Kluger, PhD Student
Department of Statistics, Stanford University

Todd Alonzo, PhD
Professor of research population and public health sciences, group statistician for Children's Oncology Group, Keck School of Medicine of USC, University of Southern California

Howard H. Yang, PhD
Staff Scientist, Laboratory of Cancer Biology and Genetics, Center for Cancer Research, NCI

Dr. Kipnis asked the panelists to address the following areas of inquiry:

1. **Does the grouping of primary, secondary, and exploratory endpoints represent families? Accordingly, should adjustments be made for each family separately or across families?**

2. **How should we define power in a meaningful way? Are there definitions that could be applied to different types of studies? Why?**

Yes, endpoint groupings are different families, Dr. Unger agreed, and primary analyses should be adjusted for multiplicity. However, he expressed ambivalence toward adjusting secondary outcomes. For trial designers, he saw advantages in focusing on meaningful hypotheses, including those having support in the literature, instead of "stretching" secondaries into an endless list.

Because the concept of loss (as advocated by Dr. Greenland) could embrace a lost idea, Dr. Unger felt exploratory ideas should not be constrained. Finally, he answered his own question from Day 1, saying he has faith in the research community to figure "things" out, whether designing a study or reading a publication. Their acumen is the preferable alternative to having a top-down authority impose some rule.

Dr. Kipnis reiterated that false results will occur, so multiplicity testing is not a constraint but a tool to protect inference using statistics. It's easier to explain results after the experiment is completed, rather than before. If an inference is not protected, others will jump on it and likely waste time and money. Conversely, researchers could fail to study something potentially valuable because early information was lost.

Dr. Dodd acknowledged that researchers undertaking multiple tests should think about which and how many hypotheses must be rejected before considering something effective; however, it is more efficient to ask two questions in one study than to conduct two separate studies. He also stated being struck by Dr. Greenland's recommendation to put observations out there as long as they are sufficiently described. He felt a balance should be struck between the strict application of adjustments to primaries

and some allowance for descriptive results for secondaries, while protecting the inferences. He added a caveat: Secondary endpoints may help show that a true effect exists when primary endpoints fail to demonstrate this.

Dr. Vickers cautioned that while p-values and CIs have their place, they remain abstract concepts, removed from the real world, so statistics alone can't provide the answers researchers are seeking. Ultimately, you can't legislate good scientific judgment and then abandon it just because you have an adjusted p-value, he explained.

Dr. Miratrix disagreed, citing an adjusted p-value as the "measure of surprise" for multiple null hypotheses that are tossed into different buckets rather than one.

Dr. Vickers noted that researchers know studies are expensive, so they will review all results. He also pointed out that significant results don't mean an effect exists; instead, they indicate something is likely to have an effect. He reiterated that experiments must be repeated.

The colleagues agreed to discuss this topic further, using real-world examples to illustrate whether multiple testing should be done, or would help, as opposed to readers applying their sound scientific judgment.

Dr. Unger sided with Dr. Vickers, saying that searching for examples in the literature would only harm the current discussion. Although excitement over a finding with a p-value under 0.05 can lead to spurious research, there are also many unobserved "examples that were not pursued due to an overreliance on multiplicity corrections."

Mr. Kluger reminded the panel there is consensus around the use of multiplicity corrections, including controlling for FDR, when testing high-dimension data for an association between a gene and a phenotype. He pointed to these adjusted tests as the preliminary examination before advancing to expensive CRISPR software for further analysis.

Dr. Izmirlian introduced interesting theoretical work on multiple testing by Mr. Kluger, whose thesis work has examined the impact of correlated tests on the error rates of omics studies, using the Benjamini–Hochberg-FDR procedure. Mr. Kluger expounded on a finding—introduced to the workshop by Dr. Miratrix—in which strong long-range, test statistic dependency creates either a few or a flood (e.g., < 50%) of false discoveries, even though the rate is controlled at < 0.1.

Dr. Izmirlian displayed violin plots from simulated data to demonstrate the usefulness of the FDR and the average power as summaries of the distribution of their per-experiment values. The FDR, which is used to protect inferences against multiplicity, is the population mean of a per-experiment unmeasured proportion, or the false discovery proportion (FDP). This is the portion of discoveries that are false. The average power can be thought of as a multiplicity-corrected power, which is used to calculate sample sizes and is the population mean of a per-experiment unmeasured proportion, or the true positive proportion (TPP). This is the portion of true signals that are discovered.

Even though FDR and average power are intended to control the distribution of their per-experiment values, Dr. Izmirlian explained, they do so poorly when the FDR procedure and average power are used in MTPs with less than 1,000 hypothesis tests.

Simulated under independence, Dr. Izmirlian's violin plots showed that only when the number of test statistics is large (e.g., 10,000 or more) do FDR and average power adequately summarize their respective distributions. This means that control of validity, using FDR, and calculation of sample size, using average power, do not guarantee much on a per-experiment basis. A second set of plots displaying data simulated under block correlation showed that the problem is exasperated when there is dependence. He mentioned that there are two procedures that directly control the probability of whether FDP exceeds a threshold, the so-called FDX and a definition of power that is based on the probability of the TPP exceeding a threshold, or the so-called TPX power. In their studies, Lehmann and Romano (2008) and Izmirlian (forthcoming) provide procedures for controlling the probability of an FDX. Dr. Izmirlian also mentioned that his forthcoming paper shows how to derive multiplicity-adjusted power, based on the distribution of the TPP.

Dr. Yang agreed that correlation decreases with increase in group size.

Returning to the overall conversation, Dr. Alonzo related his experience from designing pediatric cancer trials on making family-wise error corrections for secondary hypotheses. Although a trial's primary endpoint may be event-free survival, he emphasized each secondary stands on its own as an important comparison, ranging from examining an adverse effect to a broader topic, like quality of life. Therefore, his team does not consider all secondaries (typically 5 to 10) as part of one family. Using an example of cardiac toxicity (testing dosage level), he relayed his concern for losing a critical observation because all secondaries are adjusted as a family. However, he feels a family-wise error adjustment is better applied to quality of life, a factor composed of multiple domains.

Dr. Dodd countered, from the funding institution's perspective. He said that a researcher may miss some things when controlling a study more broadly, but magic can't always be expected to happen, like with penicillin. The risk lies in allowing researchers to look at too many things, with different controls on the validity of each finding.

Dr. Kipnis conceded that NCI, as a funding agency, pushes researchers to solve all kinds of problems (with finite resources), so the Institute is partially responsible for the large number of multiple tests.

Dr. Yang was asked why researchers perform multiplicity corrections. He answered: to prioritize, select, and publish results. Researchers know reproducibility is important, he explained, and they want to look back at prior results if the current p-value is marginal. To develop our research, we must accumulate evidence, cross-check with other laboratories, and see results from different sources, Dr. Yang said. Although statisticians don't use generalization often in the machine-learning world, the public finds it important to know the "answers."

Dr. Kipnis made what could be considered a culminating statement: Although statistics can protect inference, they will never be the definitive answer.

Workshop Recap and Next Steps

Vance Berger, PhD
Grant Izmirlian, PhD
Victor Kipnis, PhD

Concluding Remarks

The three moderators invited participants to share email addresses in the chat to continue the discussion beyond the panel and urged participants to take advantage of the shared emails.

Because it was clear that participants had more to say, Dr. Berger upheld the inaugural multiplicity workshop as a stepping stone—the beginning and not the end. He urged participants to foster collaboration, even if viewpoints differ, because sometimes these are the best collaborations.

Dr. Kipnis looked forward to a repeat edition of the current workshop, with more time for discussion allowed. Acknowledging the important conversations on the difficult topic of multiplicity, including some heated discussions, he stressed their importance. The group's discussion proved that despite breakthroughs, developing techniques is not enough. Many questions remained, he said, including how to combine techniques, apply those techniques, and work with substantive scientists who understand the biology that moves research forward.

Attendee Poll Results

The following are key findings from a poll taken of 70+ responding participants:

- Respondents mostly perform their statistics work themselves rather than collaborate with others.
- Industry sectors were represented in the following order, descending from the largest number: government, academia, nonprofit, and private industry.
- The largest group of participants reported being established in their fields. In an accompanying question, they reported having more than 11 years of experience.
- When asked about areas of expertise, participants were permitted to check all that applied. The results showed good representation in many fields. Participant responses were ranked in order, from most chosen to least: statistics, cancer, clinical trials, data science, biology, and other.

Dr. Izmirlian recommended a smaller follow-up workshop focused on tutorials for those interested in learning more about how to apply the available software packages to their statistical endeavors.