Executive Summary

July 16, 2021 Research Base Statistical Leads Meeting:

A meeting of statistical leads from the seven NCI Community Oncology Research Program (NCORP) Research Bases with statisticians from the Division of Cancer Prevention (DCP) Biometry Research Group took place on July 16, 2021.

Meeting Goals and Content:

The meeting had a few key objectives:

1) To bring together the Research Base (RB) statisticians and Discussion Leads and the DCP statisticians from the Biometry Research Group (BRG).
2) To facilitate a meeting and interchange between the two groups of statisticians.
3) To engage both sets of statisticians in discussion of statistical issues that have emerged during reviews of concepts and protocols submitted to the NCORP program.

To accomplish these objectives, the lead statistician at each of the seven Research Bases was asked to submit the statistical issues that they and their research group colleagues felt had emerged as most critical to the concept and protocol review process. Multiple concerns were submitted. The NCI staff condensed these issues into seven key questions, as follows:

1. How do Research Bases address multiplicity and multiple comparison issues in concepts and protocols?
    a. How do RBs deal with Type I error and power?
    b. How much statistical rigor should be used for secondary and exploratory endpoints?
2. How should RB statisticians distinguish between concepts and protocols in relation to their framing of primary and secondary endpoints (e.g., QOL)?
3. How do RBs statistically approach longitudinally repeated assessments in analyzing study endpoints?
4. How do RB statisticians design studies to address sub-populations, particularly underrepresented populations (e.g., gender, race, ethnicity, age)?
5. How do RB statisticians adjust for primary endpoints and secondary endpoints in CCDR studies where the nature of the intervention is premised under a multi-level framework (i.e., patient, physician, institutional level)?
6. In studies where neurocognitive function (NCF) and QOL are being concurrently evaluated, should they be adjusted for multiplicity?
7. How do statisticians address the use of historical controls in single arm trials?

The lead statistician from each of the seven Research Bases then selected a question for which he/she would conduct the meeting discussion, serving as the Question Statistical Lead (QSL). In addition, all RB lead statisticians were invited to respond to all seven questions, providing their RB's statistical perspectives on these six other questions. For each question, the QSL then developed a slide based on his/her individual response to the assigned question and in a second slide summarized the responses of the other RB statisticians to that question.

Meeting Procedures:

Following introductions, the agenda of the Research Base Statistical Leads Meeting progressed through seven sessions, each dedicated to an in-depth discussion addressing one of the seven questions and led by the assigned QSL. Following the statement of the Question and the articulation of the RB's response to the question, the

QSL summarized the responses offered by the other RB lead statisticians. Perspectives were then solicited from the statisticians representing the Division of Cancer Prevention (DCP) Biometry Research Group (BRG), who provided their views on the Question. Each question-oriented session concluded with an extended open discussion among the statisticians, both from the RBs and BRG, and other attendees.

Both RB and DCP Biometry statisticians proved to be highly engaged in this interchange. The RB statisticians present at the meeting were largely in agreement on most issues. Disagreement generally reflected the different missions of the individual RBs, as for example the nature of clinical trials and challenges faced in children by COG researchers in contrast to the adult cancer RB trials.

Summaries of the discussions addressing each of the seven statistical issues are presented below:

Statistical Question #1: How do Research Bases address multiplicity and multiple comparison issues in concepts and protocols?

General agreement existed among the RBs that the sample size/power considerations should be driven by the primary endpoint analysis. Secondary endpoints may not drive the sample size but should be protected against false (positive and negative) discoveries. Options to better control Type I error for secondary endpoints may include: 1) setting a fixed alpha level that is lower than the standard alpha=.05 (i.e., alpha=.01 regardless of the number of secondary endpoints); 2) applying a formal adjustment for multiple comparisons, such as Bonferroni; and/or 3) explicitly stipulating that any positive findings among secondary endpoints would be considered hypothesis-generating, requiring confirmation in independent study. Such strategies should address growing concerns at NIH about irreproducibility in studies in general. Justification for the level of statistical rigor applied to the secondary endpoints should be provided.

Statistical Question #2: How should RB statisticians distinguish between concepts and protocols in relation to their framing of primary and secondary endpoints (e.g., QOL)?

The concept should include a list of all primary endpoints, the sample size calculation with the justification for the applied type I error and the power. In comparison, the protocol should provide a full description of the study design with a detailed analytical plan for all endpoints, including primary, secondary, and exploratory endpoints. There was a general agreement that primary endpoint(s) should drive the power and sample size calculations. The inclusion of all relevant secondary endpoints (with estimated power) was recommended in the concept, but the list of secondary endpoints does not need to be finalized until the protocol becomes finalized.

Statistical Question #3: How do RBs statistically approach longitudinally repeated assessments in analyzing study endpoints?

Conceptually, all RBs generally agreed that longitudinal analyses are more powerful and require no multiplicity adjustment. However, RBs prefer to use a pre-specified time point, informed by discussions with their clinical colleagues regarding where the treatment effect may be most clinically meaningfully defined, while also balancing considerations of potential dropouts or problems with survival. Adjustments are made for baseline measures and stratification factors as covariates. This strategy is believed to be most interpretable for clinicians and more readily aligns with estimates of power at a single timepoint. However, if the treatment effect is generally constant over the duration of follow-up, this approach may have lower power relative to the longitudinal analysis. In general, the longitudinal data analysis is also of interest to address potentially time-dependent informative dropout as well as within subject trajectory versus between group differences.

Statistical Question #4: How do RB statisticians design studies to address sub-populations, particularly underrepresented populations (e.g., gender, race, ethnicity, age)?

Most RBs try to address sub-populations through stratified randomization, using an enriched design with efforts to encourage all eligible patients to participate. However, since trials are designed to achieve adequate power for the entire set of patients who are enrolled, subgroup analyses are generally underpowered. Thus, the focus of any subgroup analysis should be on the magnitude and direction of the treatment effect and the confidence interval and the reported findings should be transparent about what power is available for the subgroup analysis. Interaction analyses can be used to rule out potentially large differences in treatment effects between patient groups, or conversely, to establish the hypothesis that such differences may exist, which may be examined in future trials. However, interaction analyses will also typically be underpowered. An enriched design that aims to enroll a larger portion of patients from a certain sociodemographic group is sometimes used if improved representativeness from potentially *at-risk* groups is desired. The RBs also encourage the potential conduct of trials that specifically target a given group of patients identified in prior research to be at particularly high risk or in whom treatment effects are hypothesized to be different than among other patients; in such instances, the trial will be fully powered to address the primary treatment question in the subgroup. An alternative suggestion is to conduct ancillary projects in which data from multiple trials can be combined together, although this approach faces the challenge of compatibility of the data.

Statistical Question #5: How do RB statisticians adjust for primary endpoints and secondary endpoints in CCDR studies where the nature of the intervention is premised under a multi-level framework (i.e. patient, physician, institutional level)?
CCDR studies are often cluster randomized designs and analysis and sample size calculations incorporate correlation between individuals within a cluster. Similar to non-CCDR studies, key secondary objectives are identified and power calculations are performed to identify effect sizes. However, primary objectives drive the overall design and sample size and Intraclass Correlation Coefficient (ICC), which is used to estimate correlation between Individuals within a Cluster. Some of the challenges are related to selecting sites with any stratification or matching and what to do with poor performing or overachieving sites.

Statistical Question #6: In studies where neurocognitive function (NCF) and QOL are being concurrently evaluated, should they be adjusted for multiplicity?
RBs generally agreed to adjust for multiplicity when QoL and NCF are concurrently evaluated, but they expressed doubts regarding the rationale for adjusting. Generally, RBs would adjust for multiplicity if the NCF and QOL are co-primary endpoints, but not if one or both are secondary endpoints. The primary argument against adjusting for multiplicity is that NCF and QOL are fundamentally different scientific constructs, representing different scientific domains. As such, their evaluation in the same set of patients can be considered separate (independent) experiments, and thus multiplicity adjustment should not be needed. In contrast, DCP statisticians' perspective is that NCF and QOL are interdependent and lack of multiplicity adjustment in this space would lead to unprotected inference and irreproducible results. Yet, DCP is flexible about how much rigor is applied to type I error or power if both NCF and QOL are secondary as long as the applied rigor is justified in the protocol. DCP is open to omitting multiplicity adjustment if RBs can provide the evidence that NCF and QOL are independent. Gatekeeping methods may offer a solution to this multiplicity problem, although some RB statisticians expressed concern that the burden of data collection for QOL on patients as well as sites would not be justified if gatekeeping results in the QOL data not being analyzed. A composite score is sometimes used in which NCF is combined with a QOL tool; the resulting single score does not require multiplicity adjustment.

Statistical Question #7: How do statisticians address the use of historical controls in single arm trials?
The overall view is that randomized clinical trials are the gold standard and that inherent problems exist with single arm studies using historic control data. A general consensus exists among the RBs that historic controls are rarely used in cancer control trials within the cooperative group setting. However, historic controls can be used with caution if randomized clinical trials are not feasible. The main issues with historic controls are choosing comparable controls (e.g., cancer type, stage, treatment) and adjusting for covariates and confounders. Knowledge from historical controls is considered hypothesis generating and should only be used to support planning of future trials.

Future Plans:

The success of this process has encouraged us to consider future interactions in the form of conferences as well as collaboratively written documents. The questions formulated for this conference were limited to inherently statistical issues and concerns. However, future conferences will also address additional issues related to the statistical analyses of concepts and protocols during NCORP review of submitted Research Base studies. The outstanding issues are numerous and merit further attention, especially given their important impact on cancer research and clinical trials.

Attendees:

Hosts:

Cecilia Lee	DCP/COPTRG
Barbara Dunn	DCP/COPTRG

Attendee Statisticians:

Lead Research Base Statisticians:

Constantine Gatsonis	ECOG-ACRIN
Jennifer Le-Rademacher	Alliance
Joseph Unger and Bill Barlow	SWOG
Eva Culakova	University of Rochester
Emily Dressler	Wake Forest
Stephanie Pugh	NRG
Todd Alonzo	COG
Victor Kipnis	DCP Biometry

Research Base Statisticians and Attendees:

ECOG-ACRIN: Constantine Gatsonis, Jon Steingrimsson; Bob Gray, Jorean Sicks, Na An, Fengmin Zhao, Sandra Lee, Ju-Whei Lee, Ben Herman, Brad Snyder and Fenghai Duan

Alliance: Jennifer Le-Rademacher, Amylou Dueck, Gina Mazza, Jun He, Minji Lee, Paul Novotny, Heather Gunn

SWOG: Joseph Unger, Bill Barlow, Katherine Guthrie and Michael Le Blanc

University of Rochester: Eva Culakova and Joseph J. Guido

Wake Forest: Emily Dressler, Edward Ip; Anna Snavely, Lynne Wagner, Lingyi Lu and B. Levine

NRG: Stephanie Pugh, Jim Dignam, Greg Yothers, Danielle Enserro, Reena Cecchini, Hanna Bandos, Helen Huang and Kathryn Winter

COG: Todd Alonzo Brad Pollock, Ha Dang

DCP Biometry: Victor Kipnis, Lev Sirota, Kevin Dodd and Doug Midthune

NCI/DCP, DCCPS Attendees: Brenda Adjei, Kate Castro, Leslie Ford, Ann Geiger, Marge Good, Pam Maxwell, Worta McCaskill-Stevens, Jennifer Pak, Bernard Parker, Sandra Russo, Cynthia Whitman