Overview of Strategies to Address Multiplicity for Primary and Key Secondary Assessments in Confirmatory Clinical Trials

Gary G. Koch, Ph.D.
Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill

Trends and Innovations in Clinical Trial Statistics
Multiplicity Issues in Clinical Trials
April 22, 2014

I. Regulatory Success Requires Convincing Statistical Findings

1. Well-designed clinical trials
2. Relevant and reliable data
3. Well-planned analyses

II. Statistical Methodology Strengthens Robustness of Study Findings to Sources of…

1. Bias (addressed by randomization, masking, etc.)
2. Variability (addressed by sample size, covariance adjustment)
3. Spurious outcomes (addressed by covariance adjustment, multiplicity management)

III. Multiple Assessments Lead to Multiple Opportunities for Findings to be due to Chance, and so Need Control

1. Multiple endpoints
2. Comparisons among multiple treatment groups
3. Multiple inferential subgroups
4. Multiple interim analyses
5. Multiple types of analyses (re management of missing data, protocol deviations, etc.)

IV. The Issues for Multiple Assessments are

1. Avoidance of inflation of type 1 error from insufficient control
2. Avoidance of unsatisfactory loss of power (i.e., excessive type 2 error) to detect real treatment differences from overly burdensome control

V. Many clinical trials to compare two or more treatments have multiple endpoints

1. One criterion at several visits (e.g., patient global ratings of symptom control for a respiratory disorder at four visits during the treatment period)
2. Four dichotomous criteria for outcomes of treatment of stroke
3. Three time to event criteria for chronic obstructive pulmonary disease (COPD)
4. Seven measures of severity of rheumatoid arthritis
5. Three measures of severity of benign prostatic hyperplasia

VI. Strategies for Managing Multiple Endpoints with Strong Control of Type 1 Error at Specified One-sided Alpha

1. Hierarchical (or sequential) assessments. Requires p ≤ alpha at each successive step in hierarchy to continue. Formal testing stops at first step with p > alpha (i.e., closed testing for H_{01}, H_{02}, H_{03} as $H_{01}, H_{01} \cup H_{02}, H_{01} \cup H_{02} \cup H_{03}$ addressing all intersections)
2. Closed testing through multi-way averages (Lehmacher et al [1991])
3. Alpha propagation methods (e.g., Bonferroni-Holm)
4. Composite endpoints
5. Combinations of (1) – (4)

VII. Multi-Visit Clinical Trial for a Respiratory Disorder with Patient Global Ratings of Symptom Control During the Treatment Period

1. Multiple endpoints corresponding to patient global ratings as Excellent, Good, Fair, Poor, Terrible at four visits
2. Visits 2, 3 primary and Visits 1, 4 secondary (where discontinuation of concomitant medicines is applied at Visits 2, 3)
3. Comparisons between two treatments for 111 patients through a GEE repeated measures proportional odds model with treatments, centers, visits, and treatment*visits.

VII. (continued) Multi-Visit Clinical Trial for a Respiratory Disorder with Patient Global Ratings of Symptom Control During the Treatment Period

4. Multiplicity is managed with the following successive steps (which combine the closed testing method of Lehmacher et al [1991] with a hierarchical (sequential) closed testing method)

   - Step 1: test average odds ratio for Visits 2, 3; and if one-sided $p<0.025$, then proceed; else, stop.
   - Step 2: test Visit 2 and Visit 3 separately at one-sided 0.025; and if one-sided $p<0.025$ for both Visit 2 and Visit 3, then proceed; else, stop.
   - Step 3: test average odds ratio for visits 1, 4; and if one-sided $p<0.025$, then proceed; else, stop.
   - Step 4: separately test Visit 1 and Visit 4.

VII. (continued) Multi-Visit Clinical Trial for a Respiratory Disorder with Patient Global Ratings of Symptom Control During the Treatment Period

5. The respective p-values from (4) are as follows: $p<0.001$ for Visit 2, 3 average; $p<0.001$ for Visit 2 and $p=0.001$ for Visit 3; $p=0.007$ for Visit 1, 4 average; $p=0.027$ for Visit 1 and $p=0.011$ for Visit 4. With covariance adjustment for age, baseline, gender, all one-sided p-values < 0.006.
6. Nonparametric rank based analyses, such as those from stratified counterparts of Wilcoxon rank sum statistics, provide comparable results.

VIII. The National Institute of Neurological Disorders and Stroke t-PA Stroke Trial (Tilley et al [1996])

1. Multiple endpoints corresponding to four dichotomous outcomes from Barthel Index, Modified Rankin Scale, Glasgow Outcome Scale, National Institute of Health Stroke Scale (NIHSS)
2. The primary criterion for the comparison between the test and control treatments was the average odds ratio across the four dichotomous outcomes from a GEE repeated measures logistic regression model with treatments, outcomes, and treatments*outcomes
3. The separate dichotomous endpoints had a secondary role (without a multiplicity method)
4. For the primary criterion, two-sided p=0.008, with 1.7 as the estimated odds ratio and (1.2, 2.6) as its 0.95 confidence interval

VIII. (continued) The National Institute of Neurological Disorders and Stroke t-PA Stroke Trial

5. All of the separate dichotomous endpoints had two-sided $0.019<p<0.033$; and so the method of Lehmacher et al [1991] would have indicated statistical significance for each of them at two-sided alpha=0.05 through all of the two-way and three-way average odds ratios having two-sided $p<0.05$ (but this method was not pre-specified)
6. Linear counterparts of the previously described methods would support similar conclusions through corresponding averages of risk differences

IX. Clinical Trial for Chronic Obstructive Pulmonary Disease (COPD) with Three Time to Event Criteria (Saville and Koch [2013])

1. Multiple endpoints corresponding to time to severe status for lung function or death, time to very severe status for lung function or death, time to death
2. Comparisons between test and control treatments for an illustrative sample of 2,000 patients are based on the Wei-Lin-Weissfeld (WLW) method for parallel proportional hazards models for three times to events (with these including baseline lung function, current smoking status, age categories, gender, body mass index categories, race, and geographical region)
3. The separate results for the three times to events are as follows for HR (95% CI, two-sided p)
    a. Severe/death: HR=0.806 ((0.737, 0.882), p<0.001)
    b. Very severe/death: HR=0.884 ((0.755, 1.036), p=0.127)
    c. Death: HR=0.819 ((0.649, 1.033, p=0.092)

IX.(continued) Clinical Trial for Chronic Obstructive Pulmonary Disease (COPD) with Three Time to Event Criteria (Saville and Koch [2013])

4. From use of the estimated covariance matrix for the WLW method, the average HR (with equal weights for each time to event endpoint) is HR=0.836 with (0.743, 0.941) as 95% CI and two-sided p=0.003
5. With the method of Lehmacher et al [1991], statistical significance at alpha=0.05 would apply to severe/death since the pairwise averages of its hazard ratio with those for very severe/death and death have two-sided $p<0.05$
6. Nonparametric randomization based methods with less stringent assumptions provide comparable results

X. Composite Endpoints to Address Seven Measures of Severity for Rheumatoid Arthritis

1. For rheumatoid arthritis, a core set of seven measures for improvement includes three physician assessments (tender joint count, swollen joint count, physician global rating), three patient assessments (disability, patient pain, patient global rating), and acute phase reactant (ESR/CRP).
2. Felson et al [1995] developed the ACR 20 composite endpoint as a dichotomy with success being ≥20% improvement in both tender and swollen joint counts and at least three of five of the other measures.
3. This composite endpoint has recognition in FDA Guidance [Feb/99], and it can be the primary criterion for efficacy. Thus, it enables the management of multiplicity for the core set of seven measures for improvement via one composite endpoint.

X. (continued) Composite Endpoints to Address Seven Measures of Severity for Rheumatoid Arthritis

4. A related definition for ACR 50 is available; and so consideration of the sum (ACR 20 + ACR 50) is possible when reasonably satisfactory power applies to both. If statistical significance at one-sided alpha applies, inferential assessment of both ACR 20 and ACR 50 at alpha is possible without any other multiplicity adjustment via closed testing (Lehmacher et al [1991]).
5. For other disorders, a composite criterion can specify improvement in one or more measures without worsening in the others with or without specification of the measures for which improvement is necessary.

XI. A Priori Planned Integrated Analyses for Comparisons Between Dutasteride and Placebo for Benign Prostatic Hyperplasia (BPH)

(personal communication from T. Wilson, Glaxo-Smith-Kline)

1. Three randomized double blind clinical trials with essentially identical protocols for two years duration.
2. For each clinical trial, the change in AUA-SI symptoms was the primary endpoint for one year of treatment; prostate volume and maximum urinary flow were key secondary endpoints (with their multiplicity managed by Bonferroni-Holm alpha propagation method).
3. For the integrated studies, acute urinary retention was the primary endpoint for two years of treatment, and BPH-related surgery was the secondary endpoint.
4. Separate alphas applied to year one trials and year two.

XI. (continued) A Priori Planned Integrated Analyses for Comparisons Between Dutasteride and Placebo for Benign Prostatic Hyperplasia (BPH)

5. Blinding was maintained for information pertaining to year two at the time of year one analysis, and dissemination of year one results was restricted
6. Each of the separate clinical trials had two-sided $p<0.001$ at year one for AUA-SI symptoms, maximum urinary flow, and prostate volume
7. The integrated studies had $p<0.001$ at year two for acute urinary retention and BPH-related surgery
8. Year 1 NDA submitted in December 2000 and approved in November 2001. Year 2 NDA submitted in December 2001 and approved in October 2002

XII. Two Randomized, Double Blind, Multi-Regional One-Year Clinical Trials (with Essentially Identical Protocols) to Compare Low and High Doses of Test Treatment to Placebo for a Respiratory Disorder (Wang et al [2013])

1. One primary endpoint at end of one year (FEV1) for the separate trials
2. One key secondary endpoint for the separate trials (TDI)
3. Two other key secondary endpoints for the combined data from the two trials (SGRQ and exacerbations)
4. Multiplicity is managed by having the combined data from the two trials as primary for all endpoints and using Bonferroni method for the comparisons pertaining to each dose

XII. (continued) Two Randomized, Double Blind, Multi-Regional One-Year Clinical Trials (with Essentially Identical Protocols) to Compare Low and High Doses of Test Treatment to Placebo for a Respiratory Disorder (Wang et al [2013])

5. The steps for successive assessments are as follows:

   a. Assess FEV1 for each dose at one-sided alpha=0.0005 for the combined trials and one-sided alpha=0.025 for the separate trials; in a manner like that outlined in Maca et al [2002], this criterion controls type 1 error across the two doses at one-sided alpha=0.000625 (which is comparable to two-sided alpha=0.00125).

   b. Given that (a) is satisfied for at least one dose, apply (a) to TDI for that dose (or to both if (a) is met for both)

   c. Given that both (a) and (b) are met for at least one dose, then assess SGRQ at one-sided alpha=0.0125 for that dose for the combined trials (or to both doses if (a) and (b) are met for both doses)

   d. Given that (a), (b), (c) are met for at least one dose, apply (c) to exacerbations for that dose (or to both doses if (a), (b), (c) are met for both doses)

6. Table 1 of Wang et al indicates $p \leq 0.0002$ for FEV1 and TDI for both doses in the combined trials (and in each of the separate trials); also, $p \leq 0.0001$ for SGRQ for both doses in the combined trials (and in each of the separate trials), and $p \leq 0.0002$ for exacerbations for both doses in the combined trials (with $p \leq 0.004$ for both doses in one trial and $0.05 < p < 0.11$ for both doses in the other trial)

XIII. Strategies for Managing Comparisons Among Multiple Treatment Groups with Strong Control of Type 1 Error at Specified One-sided Alpha

1. Hierarchical (or sequential) assessments for targeted contrasts
2. Closed tests for global comparisons
3. Alpha propagation methods
4. Combinations of (1) – (3)

XIV. Clinical Trial to Compare Test Treatment (T), Reference Control Treatment (R), and Placebo (P) for Reducing Symptom Severity for a Respiratory Disorder

1. The clinical trial has two objectives:
   A. Superiority of T to P
   B. Non-inferiority of T to R via $(T - P) > \theta(R - P)$ where $0<\theta<1.0$

2. Multiplicity is managed with the following successive steps:
   Step 1: Test (T vs. P); and if one-sided $p<0.025$, proceed.
   Step 2: Test $((T - P)$ vs. $\theta(R - P))$; and if one-sided $p<0.025$, proceed.
   Step 3: Test (R vs. P) and (T vs. R) separately at one-sided alpha=0.025 since Step 1 contradicts $(T\leq R)*(R\leq P)=(T\leq R\leq P)$. If one-sided $p<0.025$ for (T vs. R), stop; but if one-sided $p>0.025$ for (T vs. R) and one-sided $p<0.025$ for (R vs. P), proceed.
   Step 4: Test $((T - P)$ vs. $\theta^*(R - P))$ for $\theta< \theta^*<1$.

XV. Study to Compare T, R, and P for Healing Duodenal Ulcer (Hypothetical)

1. About 100 patients per group with six week healing rates
    a. T: 80.8% ± 4.1%
    b. R: 74.4% ± 4.7%
    c. P: 50.7% ± 5.6%
2. T and R are both superior to P ($p<0.01$)
3. (T – P)/(R – P) has (0.80, 2.50) as 0.95 confidence interval
4. T preserves at least 80% of the differences between R and P, and so is reasonably non-inferior to R

XVI. Multiple Endpoints and Multiple Treatments: Koch et al [1993]

1. Multi-center clinical trial to compare three treatments for duodenal ulcer healing and avoidance of ulcer recurrence
    a. placebo, reference, test
    b. healing, healing and no recurrence
    c. test needs to be better than placebo for healing as well as healing with no recurrence
    d. test needs to be better than reference for healing with no recurrence
    e. reference needs to be better than placebo for healing
2. The two assessments for (c) are made first at the one-sided 0.025 significance level with the Hochberg extension of the Bonferroni method; if both have $p \leq 0.025$, then (d) is assessed at the one-sided 0.025 significance level; if it has $p \leq 0.025$, then (e) is assessed at one-sided 0.025.
3. Since one-sided $p \leq 0.025$ is of interest for all tests in (c), (d), and (e), sample size needs to be large enough to avoid excessive Type 2 error from addressing (c), (d), and (e) successively rather than separately.

XVII. Clinical Trial to Compare High (H) and Low (L) Doses of Test Treatment to Possibly Inactive Reference Control (R) for Time to Substantial Improvement of an Infection

1. The test treatment can demonstrate efficacy via superiority of H to R, L to R, H to L, or L to H.
2. Multiplicity is managed with the following three steps:

   Step 1: Global comparison to address H=L=R via test statistic with two degrees of freedom. If $p<0.05$, proceed.

   Step 2: Separate two-sided assessments for H=L, H=R, L=R since at most one can apply via Step 1. If two-sided $p<0.05$ applies to any, proceed.

   Step 3: Verify that at least one of the assessments from Step 2 with two-sided $p<0.05$ directionally demonstrate efficacy via superiority of H to R, L to R, H to L, or L to H

   A simulation study may be necessary to support strong control of type 1 error at one-sided alpha=0.025 for $H \leq R$, $L \leq R$, $L \leq H$, $H \leq L$ as its respective one-sided null hypotheses

XVIII. Clinical Trial to Compare High (H), Medium (M), and Low (L) Doses of Test Treatment to Placebo (P) for Lengthening Survival for Patients with a Fatal Disorder (Tangen and Koch [1991])

1. Multiplicity is managed with the following three steps:
   Step 1: Assessment of the contrast for (H+M-2P).
   If one-sided $p<0.025$, proceed.
   Step 2: Assessment of (H vs. P) and (M vs. P) separately at one-sided alpha=0.025. If both have one-sided $p<0.025$, proceed.
   Step 3: Assessment of (L vs. P) at one-sided alpha=0.025.

2. Logrank test specified as primary method, but Wilcoxon test is more sensitive to test treatment's reduction of early deaths

XVIII. (continued) Clinical Trial to Compare High (H), Medium (M), and Low (L) Doses of Test Treatment to Placebo (P) for Lengthening Survival for Patients with a Fatal Disorder (Tangen and Koch [1991])

3. Adjustment for informally specified covariables provides stronger results than counterparts without adjustment (mainly by variance reduction)
4. The one-sided p-values for the respective assessments are as follows:

Method	(H+M-2P)	H vs. P	M vs. P	L vs. P
Logrank Adjusted	0.007	0.013	0.022	0.126
Wilcoxon Adjusted	0.003	0.020	0.010	0.095
Logrank Unadjusted	0.025	0.046	0.044	0126
Wilcoxon Unadjusted	0.016	0.039	0.027	0.105

XIX. Strategies for Managing Comparisons Between Two Treatments for the Overall Population and One or More Subgroups that are Inferentially Primary

1. Hierarchical (or sequential) assessments (e.g., subgroup first), and if it has one-sided p<alpha, then overall population (and if it has one-sided p<alpha, evaluate whether the complement is supportive in the sense of at least a trend for which satisfactory power applies).
2. Hochberg's method with the relatively high correlation of assessments taken into account with a re-randomization method such as that in SAS MULTTEST.
3. An alpha propagation method such as that for Bonferroni-Holm (with correlation of assessments taken into account), particularly if the scope for inference includes key secondary endpoints or comparisons among multiple treatment groups.
4. An alpha spending method such as O'Brien-Fleming (or Pocock) for subgroups in a nested relationship to one another (e.g., all patients, subgroup with (2/3) of patients, sub-subgroup with (1/3) of patients).

XX. Clinical Trial to Compare Combination Treatment B=(A+C) to Each of its Components A and C for 476 Patients with a Urinary Tract Infection (Koch and Schwartz [2014])

1. Primary populations were all patients and subgroup of patients with complicated diagnosis
2. For efficacy to apply to fixed combination B=A+C, superiority to both A and C is necessary
3. Multiplicity for populations is managed with Hochberg's method (i.e., one-sided $p<0.025$ for both B vs. A and B vs. C for both populations ***or*** one-sided $p<0.0125$ for both B vs. A and B vs. C for at least one population); correlation of assessment for populations not taken into account

XX. (continued) Clinical Trial to Compare Combination Treatment B=(A+C) to Each of its Components A and C for 476 Patients with a Urinary Tract Infection (Koch and Schwartz [2014])

4. For both populations, one-sided $p \leq 0.002$ applies to both B vs. A and B vs. C, and so efficacy of B has demonstration for both populations
5. Discussion in Koch and Schwartz [2014] indicates that one-sided $p=0.326$ for B vs. A and one-sided $p=0.158$ for B vs. C are adequately supportive for trends for efficacy of B to apply to patients with an uncomplicated diagnosis relative to the available power
6. On the basis of (4) and (5), efficacy reasonably applies to the overall population (even though it is mainly evident for patients with a complicated diagnosis).

XXI. Vorapaxar in Secondary Prevention of Atherothrombosis

(FDA Advisory Committee Meeting, January 15, 2014)

1. Double blind placebo controlled global trial with 26449 patients in 3 strata:
    a. coronary arterial disease (CAD) as post-myocardial infarction; 17779 patients
    b. cerebrovascular disease (CVD); 4883 patients
    c. peripheral arterial disease (PAD); 3787 patients
2. Primary endpoint was time to first occurrence of CV death, MI, stroke, or urgent coronary revascularization (UCR); and key secondary endpoint was time to CV death, MI, or stroke.
3. Since another vorapaxar trial in a different population for a different indication identified unacceptable benefit risk for CVD patients, three populations became of interest; all patients, all CAD and PAD patients (excluding those with CVD history); CAD patients (excluding those with CVD history).

XXI. (continued) Vorapaxar in Secondary Prevention of Atherothrombosis (FDA Advisory Committee Meeting, January 15, 2014)

4. Both the primary endpoint and the secondary endpoint could have been primary with the Lehmacher et al (1991) method used to manage multiplicity through the Wei-Lin-Weissfeld method.
5. Multiplicity for the three populations could have been addressed formally in a sequential way with all patients first, (CAD + PAD) second, and CAD third (given one-sided $p<0.025$ for both endpoints at all previous steps). However, other strategies would be possible.
6. For both endpoints, all three populations had $p<0.025$.

XXII. Multistage Clinical Trial to Evaluate Successively More Stringent Criteria for Rare Events (Li et al [2013])

1. For new treatments to manage type 2 diabetes, a clinical trial to exclude risk of major adverse cardiac events (MACE) is potentially necessary.
2. Three primary null hypotheses are of interest for the relative risk (RR), or hazard ratio (HR), for MACE; and they are H_{01}: RR≥1.8, H_{02}: RR≥1.3, H_{03}: RR≥1.0 with H_{02} being the most primary.
3. Multiplicity for H_{01}, H_{02}, H_{03} can be addressed by their sequential assessment (with H_{01} first, H_{02} second, and H_{03} third, with contradiction of preceding hypotheses being necessary for consideration of the next hypothesis).

XXII. (continued) Multistage Clinical Trial to Evaluate Successively More Stringent Criteria for Rare Events (Li et al [2013])

4. Interim analyses are used to assess the respective hypotheses with each hypothesis having its own alpha spending function.
5. The following analysis plan for one-sided alphas has reasonably good power to contradict H_{01} and H_{02} if RR=1 as well as to contradict H_{03} if RR≤0.8.

Stage	1	2	3	4	5
Number of events	100	200	450	700	900
H_{01}	0.0125	0.0150	NA	NA	NA
H_{02}	0.00005	0.00495	0.01	0.0175	NA
H_{03}	0.00005	0.00005	0.0004	0.010	0.024

Some increase in power is provided by a Hochberg refinement for the last two stages for each hypothesis (i.e., 0.025 for H_{01}, 0.0225 for H_{02}, 0.0245 for H_{03}).

Well-planned statistical strategies enable a clinical trial to have better findings

1. Study designs with better representation of patient population, better compliance with the protocol, and sufficient sample size for study objectives
2. Better data quality through methods for reducing prevalence of missing data and protocol deviations
3. Analysis plans with covariance adjustment to increase statistical power (through reduced variance) and with multiplicity procedures to support robustness from spurious events

Presentation Abstract

Multiple assessments lead to multiple opportunities for findings to be due to chance, and so need control. These multiple assessments arise from paradigms involving multiple endpoints, comparisons among multiple treatment groups, multiple inferential subgroups, and multiple interim analyses. Issues for multiple assessments require careful attention in order to avoid inflation of type 1 error, as well as to avoid unsatisfactorily low power to detect real treatment differences from overly burdensome control.

This presentation provides an overview of some strategies to address multiplicity for primary and key secondary assessments in confirmatory clinical trials. For multiple endpoints, potential strategies include closed testing through multi-way averages, alpha propagation methods, and composite endpoints; combinations of these strategies can also be useful in some situations. Strategies to address comparisons among multiple treatment groups include targeted contrasts, alpha propagation functions, and combinations of contrasts with alpha propagation functions. For multiple inferential subgroups, alpha propagation functions are applicable, and their power can be better through accounting for correlations among assessments. For situations which have multiple inferential subgroups with additional sources of multiplicity, such as comparisons among multiple treatment groups, combinations of strategies that account for each source of multiplicity become necessary. Combinations of strategies are also needed for situations which have multiple interim analyses in addition to one or more other sources of multiplicity, such as multiple endpoints and/or multiple inferential subgroups. Several examples are provided in this discussion to illustrate the planning of strategies to address multiplicity in ways that can provide reasonable statistical power while avoiding inflation of type 1 error.

Gary G. Koch bio

Gary G. Koch, Ph.D., D.Sc. (Hon) is a Professor of Biostatistics and Director of the Biometric Consulting Laboratory at the University of North Carolina at Chapel Hill, where he has served on the faculty since 1968. He received the BS in Mathematics and the MS in Industrial Engineering from The Ohio State University, the Ph.D. in Statistics from the University of North Carolina at Chapel Hill, and D.Sc. (Honorary) from De Montfort University. His principal research interest is the development of statistical methodology for the analysis of categorical data and its corresponding applications to a wide range of settings in health sciences. He has an extensive record of publication in statistics and in collaborative work in health sciences research. The biopharmaceutical topics, which are addressed by his publications, include crossover studies, multi-center studies, longitudinal (multi-visit) studies, rank methods for ordered outcomes, covariance analysis and adverse experience data analysis. He has previously served on the editorial boards of The American Statistician, Biometrics, Drug Information Journal, and Technometrics, and he is currently a member of the editorial boards for Statistics in Medicine and The Journal of Biopharmaceutical Statistics.

References

- Alosh, M. and Huque, M. [2013]. Multiplicity considerations for subgroup analysis subject to consistency constraint. *Biometrical Journal*, 55(3): 444-462.
- Bauer, P. [1991] Multiple testing in clinical trials. *Statistics in Medicine*, 10: 871-890.
- DeMets, D.L. and Lan, K.K.G. [1994] Interim analysis: The alpha spending function approach. *Statistics in Medicine*, 13(13/14): 1341-1352.
- Denne, J.S. and Koch, G.G. [2001] Monitoring a clinical trial with multiple hypotheses concerning the treatment effect on a single primary endpoint. *Statistics in Medicine*, 20: 2801-2812.
- Dmitrienko, A. and Tamhane, A.C., [2007]. Gatekeeping procedures with clinical trial application. *Pharmaceutical Statistics*, 6: 171-180.
- Dmitrienko, A., Offen, W.W., Westfall, P.H. [2003]. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine*, 22(15): 2387-2400.
- Hochberg, Y. [1988]. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75: 800-802.
- Hochberg, Y. and Tamhane, A.C. *Multiple comparison procedures*, Wiley: New York, 1987.
- Holm, S. [1979] A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6: 65-70.
- Hommel, G., Bretz, F., Maurer, W. [2011]. Multiple hypotheses testing based on ordered p-values—A historical survey with appicatoins to medical research. *Journal of Biopharmaceutical Statistics*, 21: 595-609.
- Koch, G.G. [1997] Discussion for “p-value adjustments for subgroup analyses.” *Journal of Biopharmaceutical Statistics*, 7(2): 323-331.
- Koch, G.G., Davis, S.M., Anderson, R.L. [1998]. Methodological advances and plans for improving regulatory success for confirmatory studies. *Statistics in Medicine*, 17: 1675-1690.

References (continued)

- Koch, G.G. and Schwartz, T.A. [2014]. An overview of statistical planning to address subgroups in confirmatory clinical trials. *Journal of Biopharmaceutical Statistics*, 24(1): 72-93.
- Kong, L., Koch, G., Liu, T., Wang, H. [2005]. Performance of some multiple testing procedures to compare three doses of a test drug to placebo. *Pharmaceutical Statistics*, 4: 25-35.
- Lehmacher, W., Wassmer, G., Reitmeir, P. [1991]. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*, 47: 511-521.
- Li, J. [2013]. Testing each hypothesis marginally at alpha while still controlling FWER: How and when. *Statistics in Medicine*, 32: 1730-1738.
- Li, S., Hussey, M.A., Schwartz, T.A., Koch, G.G. [2013]. A multistage analysis strategy for a clinical trial to address the successively more stringent criteria for a primary endpoint with a low event rate. *Pharmaceutical Statistics*, 12: 65-73.
- Maca, J., Gallo, P., Branson, M., Maurer, W. [2002]. Reconsidering some aspects of the two-trials paradigm. *Journal of Biopharmaceutical Statistics*, 6: 89-97.
- Marcus, R., Peritz, E., Gabriel, K.R. [1976]. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63: 655-660.
- O'Brien, P.C. [1984]. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40: 1079-1087.
- O'Brien, P.C. and Fleming, T.R. [1979]. A multiple testing procedure for clinical trials. *Biometrics*, 35: 549-556.
- Sarkar, S. and Chang, C.K. [1997]. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92: 1601-1608.
- SAS Institute, Inc. [2012]. SAS OnlineDoc, Version 12.1 with PDF Files, Chapter 61, The MULTTEST Procedure. SAS Institute, Inc.: Cary, NC.

References (continued)

- Saville, B.R. and Koch, G.G. [2013]. Estimating covariate-adjusted log hazard ratios in randomized clinical trials using Cox proportional hazards models and nonparametric randomization based analysis of covariance. *Journal of Biopharmaceutical Statistics*, 23: 477-490.
- Shaffer, J. [1986]. Modified sequentially rejective multiple test problems. *Journal of the American Statistical Association*, 81: 826-831.
- Somerville, M., Wilson, T., Koch, G., Westfall, P. [2005]. Evaluation of a weighted multiple comparison procedure. *Pharmaceutical Statistics*, 4: 7-13.
- Tangen, C.M. and Koch, G.G. [2001]. Non-parametric analysis of covariance for confirmatory randomized clinical trials to evaluate dose-response relationships. *Statistics in Medicine*, 20: 2585-2607.
- Tilley, B.C., Marler, J., Geller, N.L., Lu, M., Legler, J., Brott, T., Lyden, P., Grotta, J. [1996]. Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial. *Stroke*, 72: 2136-2142.
- Troendle, J.F. and Legler, J.M. [1998]. A comparison of one-sided methods to identify significant individual outcomes in a multiple outcome setting: Stepwise tests or global tests with closed testing. *Statistics in Medicine*, 17: 1245-1260.
- Wang, S., Bretz, F., Dmitrienko, A., Hsu, J., Hung, H.M.J., Huque, M., Koch, G. [2013]. Panel forum on multiple comparison procedures: A commentary from a complex trial design and analysis plan. *Biometrical Journal,* 55(3): 275-293.
- Wei, L.J., Lin, D.Y., Weissfeld, L. [1989]. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of American Statistical Association*, 84: 1065-1073.
- Westfall, P.H., Tobias, R.D., Rom, D., Wolfinger, R.D., Hochberg, Y. *Multiple Comparisons and Multiple Tests using the SAS System.* SAS Institute, Inc.: Cary, NC, 1999.