American Statistical Association
2004 FDA/Industry Statistics Workshop
September 23, 2004

"Discussion for 'Gate-Keeping Strategy Session'"

Gary G. Koch, Ph.D.
Department of Biostatistics, School of Public Health
University of North Carolina at Chapel Hill
Chapel Hill, N.C., 27599-7420

Multiple assessments lead to multiple opportunities for findings to be due to chance and so need control (Koch and Gansky [1996])

1. Multiple endpoints (address through composites as global criteria or closed procedures such as Hommel, Holm, Hochberg, or Hailperin-Ruger extensions of Bonferroni method, or Westfall and Young re-sampling method)
2. Multiple treatment comparisons (address through sequential contrasts as global criteria or closed procedures)
3. Multiple subgroups (address as descriptive, or when inferential, with closed procedure or significance level spending function)
4. Multiple interim analyses (address through significance level spending function such as O'Brien-Fleming method)

The issues for multiple assessments are to avoid loss of validity because of inflation of Type I error from insufficient control and loss of power (i.e., excessive Type II error) from over-control.

Basic methods for two or more primary response variables

1. Hierarchical or step down assessment. Require $p \leq \alpha$ at each step in hierarchy. Stop testing at first step with $p > \alpha$.
2. O'Brien [1984, *Biometrics*, *40*, 1079-87] nonparametric rank sum method; rank each endpoint across all patients, add ranks across endpoints for each patient, compute counterpart of Kruskal-Wallis statistic (via SAS PROC FREQ, CMH2 with table scores and no print). Addresses global null hypothesis of no differences among treatments for all endpoints versus alternative of consistent pattern of favorable differences for most
   a. can account for strata in study design
   b. can incorporate nonparametric covariance adjustment
   c. can provide assessments for specific endpoints with control for overall Type I error through extension by Lehmacher et al [1991, *Biometrics*, *47*, 511-532]
3. Hochberg [1988, *Biometrika*, *75*, 800-802] extension of Bonferroni method. For c non negatively correlated criteria, it requires $p \leq \alpha$ for all $c$, or $p \leq (\alpha/2)$ for $(c-1)$, or $p \leq (\alpha/3)$ for $(c-2)$, ... or $p \leq (\alpha/c)$ for 1. It is more useful than (2) when one or more endpoints could have no difference among treatments in an unknown pattern; (also see Westfall et al [1999])

The O'Brien method can be particularly useful for two or more dichotomous endpoints

1. The sum of indicator variables (rather than ranks) produces an ordered categorical variable (e.g., 0, 1, 2) which usually provides more power than a dichotomy
   a. a moderate and rigorous criterion for onset and duration of relief for tension headache (e.g., onset in 1 hour and duration for 8 hours, onset in 0.5 hour and duration for 24 hours)
   b. a more inclusive and a more critical set of cardiovascular events
   c. two or more dichotomous criteria for favorable outcome relative to stroke
   d. number of time intervals with favorable outcome during a follow-up period
2. Dichotomies can alternatively be integrated on the logistic scale with methods for generalized estimating equations (GEE) being one potential strategy (see Tilley, et al [1996, *Stroke*, *27*, 2136-42])

Tilley, B.C. et al [1996, Stroke, 27, 2136-42, "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 (New England Journal of Medicine, 333 (24), December 14, 1995, 1581 87)]

1. dichotomous outcomes from the Barthel Index, Modified Rankin Scale, Glasgow Outcome Scale and National Institutes of Health Stroke Scale (NIHSS) were integrated as a composite endpoint through generalized estimating equations (GEE) for logistic regression
2. the overall result had p=0.008 for 1.7 as the estimated odds ratio and (1.2, 2.6) as its 0.95 confidence interval
3. the respective components had $0.019 < p < 0.033$, although a method to address their multiplicity (e.g., Lehmacher et al [1991]) was not specified
4. global test was considered helpful for interpretation in a setting where no single measure is accepted and where evidence of efficacy should be a "consistent and persuasive" difference between treatments
5. tests of homogeneity among components are possible with GEE through treatment $\times$ component interaction

Closed testing procedures (Bauer [1991]). All hypotheses in a specified set can have assessment at level $\alpha$ if rejection of a specific hypothesis must require rejection of that hypothesis at level $\alpha$ and rejection of all possible intersections of that hypothesis with all other hypotheses at level $\alpha$.

1. Step down assessment (e.g., $H_{01}$, $H_{02}, H_{03}$ as $H_{01}, H_{01} \cup H_{02}, H_{01} \cup H_{02} \cup H_{03}$ for which intersections with $H_{01}$ correspond to $H_{01}$ and for which intersections with $H_{01} \cup H_{02}$ correspond to $H_{01} \cup H_{02}$)
2. Bonferroni Holm
3. Lehmacher et al [1991] for multiple endpoints
4. All multi-group comparisons that contain a specified comparison between two groups relative to the comparison of those two groups (e.g., with three groups, rejection of both $\mu_1 = \mu_2 = \mu_3$ and $\mu_1 = \mu_2$ at level $\alpha$ is necessary for rejection of $\mu_1 = \mu_2$ relative to experimentwise $\alpha$ for $\{\mu_1 = \mu_2, \mu_1 = \mu_3, \mu_2 = \mu_3\}$).

Some studies have two (or more) primary hypotheses as the first objective and one (or more) additional primary (or secondary) hypotheses as the second objective

1. A study to compare test and control treatments for two primary endpoints as the first objective and for one key secondary endpoint as the second objective
2. A study to compare high dose and low dose to control for one primary endpoint as the first objective and for one key secondary endpoint as the second objective
3. A study to demonstrate non-inferiority of high dose and low dose to active reference control as the first objective and to demonstrate superiority of high dose and low dose to active reference control as the second objective

Closed testing procedures to control the experimentwise type I error at $\alpha$ for studies with two (or more) primary hypotheses as the first objective and one (or more) additional primary (or secondary) hypotheses can be complex (Dmitrienko et al [2003])

1. Let $H_{01}$ and $H_{02}$ denote two hypotheses for the first objective and $H_{03}$ denote the hypothesis for the second objective.
2. One strategy is to apply the Hochberg method to $H_{01}$ and $H_{02}$ first, and only if both have $p < \alpha$ is $H_{03}$ tested at $\alpha$.
   a. Its structure has $H_{01}$, $H_{02}$, and $H_{03}^{*} = H_{01} \bigcup H_{02} \bigcup H_{03}$ as the real hypotheses.
   b. For rejection of $H_{01}$, rejection of $H_{01}, H_{01} \bigcap H_{02}, H_{01} \bigcap H_{03}^{*} = H_{01}$, $H_{01} \bigcap H_{02} \bigcap H_{03}^{*} = H_{01} \bigcap H_{02}$ is required and is addressed by Hochberg method for $H_{01}$ and $H_{02}$.
   c. For rejection of $H_{02}$, rejection of $H_{02}$ and $H_{01} \bigcap H_{02}$ by Hochberg method is sufficient by process like (b)
   d. For rejection of $H_{03}$, rejection is needed for $H_{01}$, $H_{02}$, and $H_{03}$
3. A second strategy is to assess $H_{01}$ and $H_{02}$ first in a way which enables assessment of $H_{03}$ second if either $H_{01}$ or $H_{02}$ is rejected by Hochberg method
   a. Its structure has $H_{01}$, $H_{02}$, and $H_{03}^{*} = (H_{01} \bigcap H_{02}) \bigcup H_{03}$ as the real hypotheses
   b. For rejection of $H_{01}$, rejection of $H_{01}, H_{01} \bigcap H_{02}, H_{01} \bigcap H_{03}^{*} =$ $[(H_{01} \bigcap H_{02}) \bigcup (H_{01} \bigcap H_{03})]$, and $H_{01} \bigcap H_{02} \bigcap H_{03}^{*} = H_{01} \bigcap H_{02}$ is required; since $p < \alpha$ is necessary for $H_{01}$, $H_{02}$, and $H_{03}$ or $p < (\alpha/2)$ is necessary for $H_{01}$ if $p > \alpha$ for $H_{02}$ or $H_{03}$ (via the Hochberg method for $H_{01} \bigcap H_{02}$ and $H_{01} \bigcap H_{03}$), the result for $H_{03}$ affects the stringency for the test of $H_{01}$ as well as that for $H_{02}$.
   c. The considerations for rejection of $H_{02}$ are like those for the rejection of $H_{01}$.
   d. For rejection of $H_{03}$, $p < \alpha$ is necessary for $H_{01}$, $H_{02}$, and $H_{03}$ or $p < (\alpha/2)$ is necessary for $H_{03}$ and $H_{01}$ or $H_{02}$.
4. The strategies become more complex when the number of hypotheses for the first objective is $\geq 3$ or the number of hypotheses for the second objective is $\geq 2$

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 at the 0.05 significance level with the Hochberg method; if both have $p \leq 0.05$, then (d) is assessed at the 0.05 significance level; if it has $p \leq 0.05$, then (e) is assessed at the 0.05 significance level.
3. Since $p \leq 0.05$ is of interest for all tests in (c), (d), and (e), sample size needs to be large enough to avoid excessive Type II error from addressing (c), (d), (e) successively rather than separately

For a study with high dose (H), low dose (L), and active reference control (R), let $H_{0H1}$ and $H_{0L1}$ denote hypotheses to demonstrate non-inferiority for H and L relative to R and let $H_{0H2}$ and $H_{0L2}$ denote corresponding hypotheses to demonstrate superiority.

1. The usual strategy is to evaluate non-inferiority for both doses as the first objective with the Hochberg method
2. If both doses have $p < \alpha$ for non-inferiority, then superiority for both doses can be evaluated as the second objective with the Hochberg method in a closed test
3. If superiority for a dose is of interest when only one dose demonstrates non-inferiority, the assessment of both non-inferiority and superiority becomes more complex
    a. the actual hypotheses are $H_{0H1}$, $H_{0L1}$, $(H_{0H1} \bigcap H_{0L1}) \bigcup H_{0H1} \bigcup H_{0H2} = H_{0H2}$ and $(H_{0H1} \bigcap H_{0L1}) \bigcup H_{0L1} \bigcup H_{0L2} = H_{0L2}$
    b. With the Hochberg method for $H_{0H1}$ and $H_{0L1}$ (as well as other subsets of hypotheses), and closed testing, the rejection of $H_{0H1}$ requires rejection of $H_{0H1}$, $H_{0H1} \bigcap H_{0L1}$, $H_{0H1} \bigcap H_{0H2} = H_{0H1}$, $H_{0H1} \bigcap H_{0L2}$; also, all three-way and four-way intersections involving $H_{0H1}$ are hypotheses like the preceding ones. Thus, for closed testing, rejection of $H_{0H1}$ requires $p < \alpha$ for all the hypotheses $H_{0H1}$, $H_{0L1}$, and $H_{0L2}$ or $p < (\alpha/2)$ for $H_{0H1}$ if $p > \alpha$ for $H_{0L1}$ or $H_{0L2}$; in this way, the result for superiority concerning low dose affects the stringency of the assessment of non-inferiority for $H_{0H1}$

Statistical considerations for composite endpoints (i.e., an integration of separate endpoints)

1. the statistical role of a composite endpoint
    a. improvement of statistical power (by use of more information)
    b. management of multiple comparisons across multiple endpoints
    c. more comprehensive scope
2. issues for interpretation
    a. self-standing validity versus bridge to components
    b. homogeneity of components (and implications to power)
    c. criteria for non-inferiority, particularly for components

Felson, et al [1995, Arthritis and Rheumatism, 38, The American College of Rheumatology Preliminary Definition of Improvement in Rheumatoid Arthritis]

1. this paper developed the ACR20 composite endpoint for rheumatoid arthritis
   a. at least 20% improvement from baseline in both tender and swollen joint counts and at least 20% improvement in 3 measures among patient global assessment, physician global assessment, patient pain, disability, and acute phase reactant (ESR or CRP)
   b. it is in FDA Guidance for Industry: Clinical Development Programs for Drugs Devices, and Biological Products for the Treatment of Rheumatoid Arthritis (Feb/99)
2. this endpoint is considered interpretable in a self-standing way and can be primary
3. multiplicity among the components and the random variation of treatment effects across them can make their role more supportive than inferential; Pincus et al [2003] describe
   a. "Patient Only" counterpart to ACR20 (2 of 3 among patient global assessment, patient pain, and disability have $\geq$ 20% improvement)
   b. "Physician Only" counterpart to ACR20 (2 of 3 among tender joints, swollen joints, and physician global assessment have $\geq$ 20% improvement)

   Given statistical significance for ACR20, the assessment of the "Patient Only" endpoint and the "Physician Only" endpoint can have multiplicity addressed with the Hochberg method
4. a related definition for ACR50 is available; and so consideration can be given to the sum of ACR20 and ACR50 as a composite variable when a consistent pattern for both at least 20% and at least 50% improvement is expected and greater weight for ACR50 than ACR20 is of interest; also given its statistical significance, separate assessment of ACR20 and ACR50 can be done without any multiplicity adjustment

Two NIH sponsored studies in psychiatry have time to treatment discontinuation for any reason as their primary endpoint (Davis, Koch, Davis, LaVange [2003])

1. A study to compare olanzapine, quetiapine, risperidone, ziprasidone, and perphenazine for about 1500 patients with schizophrenia; supportive endpoints are
   a. time as doing well (CGI-Severity $\leq$ 3 or CGI-Severity = 4 and CGI-Severity Change from baseline $\geq$ 2) during follow-up
   b. composite ordering (completion of follow-up and doing well, discontinuation for non-informative reason and doing well, completion of follow-up and not doing well, discontinuation for non-informative reason and not doing well, discontinuation due to adverse event)
2. A study to compare olanzapine, quetiapine, risperidone, and placebo for about 400 patients with Alzheimer's Disease; supportive endpoints are
   a. time for which CGI-C is minimally improved or better
   b. composite ordering (completion with improvement, discontinued with improvement, completed without improvement, discontinued without improvement, discontinued for adverse event)
   c. CGI-C at 12 weeks as minimally improved or better

A priori planned integrated analyses for dutasteride as treatment for benign prostatic hyperplasia (BPH) (Personal communication from T.Wilson, Glaxo-Smith-Kline)

1. Three primary phase III studies (two in the US and one international); each was randomized, placebo-controlled, double blind, 2 year duration; they had identical inclusion/exclusion criteria, equal sample sizes for dutasteride and placebo, and 4 week placebo run-in period.
2. For each study separately, the change in AUA-SI symptoms was the primary endpoint for 1 year of treatment; prostate volume and maximum flow were key secondary endpoints (with multiplicity managed by Bonferroni-Holm)
3. For the combined studies, acute urinary retention was the primary endpoint for 2 years of treatment with BPH related surgical intervention by year 2 as the key secondary endpoint
4. Blinding was maintained for information pertaining to year 2 at time of year 1 analysis and dissemination of year 1 results was restricted so as not to have adverse influence on year 2 data
5. Each study separately had $p < 0.001$ at year 1 for AUA-SI symptoms, urinary flow, prostate volume
6. The combined studies had $p < 0.001$ at year 2 for acute urinary retention and BPH related surgical intervention.
7. Year 1 NDA submitted in December 2000 and approved in November 2001. Year 2 NDA submitted in December 2001 and approved in October 2002.

Ways of managing missing data to evaluate robustness of results

1. replace missing values by nearest preceding observed value (i.e., last observation carried forward (LOCF)) or corresponding rank
2. replace missing values by worst possible value
3. replace missing values by best possible value
4. maintain missing as missing and only analyze actually observed values
5. replace missing values by estimates from a "reasonable statistical model"
6. replace missing values for control by a good value and those for test medicine with a bad value

There is no "clearly correct" method for managing missing data. The only feasible goal for methods such as (1)-(6) is to support agreement of findings across them, and this is more difficult to achieve when the prevalence of missing data is higher and/or findings are weaker. Otherwise, power is decreased (or Type II error is increased) in the sense that rejection of null hypothesis under each of two or more methods for missing data has lower probability than such rejection for only one such pre-specified method.

Well-planned statistical strategies enable a clinical trial to have convincing 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 for enhanced reliability
3. analysis plans with covariance adjustment to increase statistical power (through reduced variance) and with multiplicity procedures to support robustness from spurious events

## References

Bauer, P. [1991]. Multiple testing in clinical trials. Statistics in Medicine, 10: 871-890.

Davis, S.M., Koch, G.G., Davis, C.E., and LaVange, L.M. [2003]. Statistical approaches to effectiveness measurement and outcome-driven re-randomizations in the clinical antipsychotic trials of intervention effectiveness (CATIE) studies. Schizophrenia Bulletin, 29(1): 73-80.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. [1994]. Analysis of Longitudinal Data. Oxford: Oxford University Press.

Dmitrienko, A., Offen, W.W., and Westfall, P.H. [2003]. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics in Medicine, 22: 2387-2400.

Felson, D.T., Anderson, J.J., Boers, M., Bombardier, C., Furst, D., Goldsmith, C., Katz, L.M., Lightfoot, R. Jr., Paulus, H., Strand, V., et. al. [1995]. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis and Rheumatism, 38: 727-735.

Friedman, L.M, Furberg, C.D., and DeMets, D.L. [1998]. Fundamentals of Clinical Trials, Springer-Verlag Inc., New York.

Hochberg, Y. [1988]. A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75: 800-802.

ICH-E9 Expert Working Group. [1999]. ICH harmonized tripartite guideline: statistical principles for clinical trials. Statistics in Medicine, 18:1905-1942.

Koch, A. and Rohmel, J. [2004]. Hypothesis testing in the gold standard design for proving the efficacy of an experimental treatment relative to placebo and a reference. Journal of Biopharmaceutical Statistics, 14(2): 315-326.

Koch, G.G. [2000]. Discussion for 'Alpha calculus in clinical trials: considerations and commentary for the new millennium'. Statistics in Medicine, 19:781-784.

Koch, G.G., Davis, S.M., and Anderson, R.L. [1998]. Methodological advances and plans for improving regulatory success for confirmatory studies. Statistics in Medicine, 17:1675-1690.

Koch, G.G. and Gansky, S.A. [1996]. Statistical considerations for multiplicity in confirmatory protocols. Drug Information Journal, 30: 523-534.

Koch, G.G. and Tangen, C.M. [1999]. Nonparametric analysis of covariance and its role in noninferiority clinical trials. Drug Information Journal, 33:1145-1160.

Lehmacher, W., Wassmer, G., and Reitmeir, P. [1991]. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics, 47:511-532.

Little, R.J.A. and Rubin, D.B. [1987]. Statistical Analysis with Missing Data. New York: John Wiley and Sons, Inc.

Moye, L.A. [2000]. Alpha calculus in clinical trials: considerations and commentary for the new millennium. Statistics in Medicine, 19:767-779.

O'Brien, P.C. [1984]. Procedures for comparing samples with multiple endpoints. Biometrics, 40: 1079-1087.

Pincus, T., Koch, G., Lei, H., et al. [2004]. Patient preference for placebo, acetaminophen (paracetamol) or celecoxib efficacy studies (PACES): two randomised, double blind, placebo controlled, crossover clinical trials in patients with knee or hip osteoarthritis. Ann Rheum Dis, 63: 931-939.

Pincus, T., Strand, V., Koch, G., Amara, I., Crawford, B., Wolfe, F., Cohen, S., and Felson, D. [2003]. An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from that of placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the disease activity score (DAS) in a rheumatoid arthritis clinical trial. Arthritis and Rheumatism, 48(3): 625-630.

Senn, S. [1997]. Statistical Issues in Drug Development. Wiley, Chichester, UK.

Senn, S. [2000]. Consensus and controversy in pharmaceutical statistics. *The Statistician*, 49: part 2, 1-22.

Tilley, B.C., Marler J., Geller, N.L., Lu, M., Legler, J., Brott, T., Lyden, P., and 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, 27:2136-2142.

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.

Westfall, P.H., Tobias, R.D., Wolfinger, R.D., and Hochberg, Y. [1999]. Multiple Comparisons and Multiple Tests Using the SAS System. SAS Institute, Inc., Cary, N.C.