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

"Non-Inferiority in Confirmatory Active Control Clinical Trials: Concepts and Statistical Methods"

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

Statistical methodology strengthens robustness of study findings to sources of

1. bias
2. spurious events
3. variability

Areas for statistical attention are

1. structure for study design
2. schedule and procedures for data collection
3. plans for primary data analyses

Placebo or active control (Koch, Davis, and Anderson [1998])

1. superiority is the objective relative to placebo, but an estimate of effect size may need to be appealing as well
2. non-inferiority (or equivalence) through a sufficiently well located and/or narrow confidence interval is the objective relative to an active control, but hypothetical superiority to placebo can be an issue; the role of the confidence interval is to show that potential inferiority to the active control is sufficiently small that
   a. superiority to placebo indirectly applies and thereby efficacy
   b. no clinical relevance applies and so efficacy is as good as the active control
3. superiority after demonstration of non-inferiority is sometimes possible

Study to compare two dosing regimens to heal duodenal ulcers by 4 weeks with ranitidine

1. Large sample size for "once per day" and "twice per day"
2. Healing rates were about
   a. once per day 72% ± 2%
   b. twice per day 78% ± 2%
3. Two-sided 95% Confidence Interval for "Once-Twice" (-11%, -1%)
4. FDA Advisory Committee could not approve equivalence
5. FDA ultimately approved "once per day" as efficacious by having healing rate well above "most optimistic" estimate of about 55% for placebo

Studies to show non-inferiority

1. Similar cure rates for anti-infective medicines
2. Similar measures for pain relief for analgesic medicines
3. Similar rates for death or myocardial infarction at specified time points after initial treatment for acute cardiovascular disorders
4. Similar patterns for overall survival during follow-up for treatment regimens in oncology trials

Alternative designs are available to address non-inferiority in confirmatory clinical trials.

1. Test (T) versus Usual Dose of Reference (R)
2. Test (T) versus Two Doses of Reference (R1, R2)
3. Two (or Three) Doses of Test versus Reference (e.g., T1, T2, R)
4. Two (or Three) Doses of Test versus Two Doses of Reference (e.g., T1, T2, R1, R2)
5. Test, Reference, Placebo in 2 : 2 : 1 randomization
6. Two (or Three) Doses of Test, Reference, Placebo

Design 1 requires more external assumptions. The other designs require more sample size and possibly methods to address multiple comparisons.

Hypothesis pertaining to superiority, $\mathrm{H}_{0:S}$: T is not superior to P versus $\mathrm{H}_{A:S}$: T is superior to P

1. $\mathrm{H}_{0:S}$: $T \leq P$ versus $\mathrm{H}_{A:S}$: $T > P$ with bigger being better; T and P could be rates or means.
2. $\mathrm{H}_{0:S}$: $(T - P) \leq 0$ versus $\mathrm{H}_{A:S}$: $(T - P) > 0$
3. $\mathrm{H}_{0:S}$: $(T/P) \leq 1$ versus $\mathrm{H}_{A:S}$: $(T/P) > 1$

Note that $\mathrm{H}_{0:S}$ and $\mathrm{H}_{A:S}$ are one-sided because superiority for T is one-sided. When a two-sided test is applied to a hypothesis concerning superiority for T, a statistically significant result only demonstrates superiority when its direction favors T; and so it is really a one-sided test. Thus, a two-sided test at the 0.05 significance level for superiority of T is really a one-sided test at the 0.025 significance level. In this sense, the criterion for efficacy of T is superiority to P at the one-sided 0.025 significance level.

Hypothesis pertaining to non-inferiority, $\mathrm{H}_{0:NI}$: T is inferior to R versus $\mathrm{H}_{A:NI}$: T is non-inferior to R

1. $\mathrm{H}_{0:NI}$: $T \leq (R - \Delta_{NI})$ versus $\mathrm{H}_{A:NI}$: $T > (R - \Delta_{NI})$ with $\Delta_{NI} > 0$ and bigger responses being better
2. $\mathrm{H}_{0:NI}$: $(T - R) \leq -\Delta_{NI}$ versus $\mathrm{H}_{A:NI}$: $(T - R) > -\Delta_{NI}$
3. $\mathrm{H}_{0:NI}$: $(T/R) \leq (R - \Delta_{NI})/R$ versus $\mathrm{H}_{A:NI}$: $(T/R) > (R - \Delta_{NI})/R$

Usually $\Delta_{NI} = (1 - L)(R - P)$ where L is the fraction of $(R - P)$ that $(T - P)$ needs to preserve; and so $(1 - L)$ is the fraction for which non-inferiority allows lack of preservation; with $\Delta_{NI} = (1 - L)(R - P)$, $(R - \Delta_{NI})/R = \{L + (1 - L)/(R/P)\}$
Note that non-inferiority is one-sided. The significance level for its demonstration is one-sided 0.025 since a major role of non-inferiority to R is to imply superiority of T to P and thereby efficacy of T.

Hypothesis pertaining to equivalence, $\mathrm{H_{0:E}}$: $\{T$ is inferior to R by being too small or T is inferior to R by being too big $\}$versus $\mathrm{H_{A:E}}$: $\{T$ is equivalent to R by being neither too small nor too big$\}$

1. $\mathrm{H_{0:E}}$: $\{T \leq (R - \Delta_{\mathrm{E}})$ or $T \geq (R + \Delta_{\mathrm{E}})\}$ versus $\mathrm{H_{A:E}}$: $\{(R - \Delta_{\mathrm{E}}) < T < (R + \Delta_{\mathrm{E}})\}$ with $\Delta_{\mathrm{E}} > 0$
2. $\mathrm{H_{0:E}}$: $|T - R| \geq \Delta_{\mathrm{E}}$ versus $\mathrm{H_{A:E}}$:$|T - R| < \Delta_{\mathrm{E}}$
3. $\mathrm{H_{0:E}}$: $\{(T/R) \leq (R - \Delta_{\mathrm{E}})/R$ or $(T/R) \geq (R + \Delta_{\mathrm{E}})/R\}$ versus $\mathrm{H_{A:E}}$: $\{(R - \Delta_{\mathrm{E}})/R < (T/R) < (R + \Delta_{\mathrm{E}})/R\}$

Equivalence is the same as non-inferiority in both the direction of not too small and the direction of not too big. Its demonstration requires two one-sided tests or a corresponding two-sided confidence interval. For purposes of efficacy, the significance level of each of the two tests is usually one-sided 0.025, or inference is based on the corresponding two-sided 0.95 confidence interval. However, for bioavailability parameters in bioequivalence studies, the one-sided 0.05 significance level and the 0.90 two-sided confidence interval are used.

The criterion that $(T - P)/(R - P) \geq L$ specifies that T needs to preserve at least $100L\%$ of the effect that R does relative to P. Its scope also includes the following considerations:

1. $\frac{T-P}{R-P} \geq L \leftrightarrow \frac{T-P}{R-P} - 1 \geq L - 1 \leftrightarrow \frac{T-R}{R-P} \geq -(1 - L)$,

   $\frac{T-R}{R-P} \geq -(1 - L) \leftrightarrow (T - R) \geq -(1 - L)(R - P)$

   i.e., $(1 - L)(R - P)$ is the bound for inferiority for which no excess must be demonstrated in order to have efficacy
2. P can be an assigned or known value such as 0 or some minimal response level; if $P = 0$, one has $\frac{T}{R} \geq L$ or $\frac{(T-R)}{R} \geq -(1 - L)$ or $(T - R) \geq -(1 - L)R$
3. $(T - R)$ can come from one study and $(R - P)$ can come from another study or an historical data base, provided that the two patient populations are comparable;

Additional considerations for confirmatory clinical trials to demonstrate non-inferiority

1. Multiplicity in endpoints and treatment comparisons (management of overall significance level and power)
2. Roles for intention-to-treat and per protocol populations and management of missing data (for non-inferiority, management of missing data should be in greater harmony with $\mathrm{H_{0:NI}}$ than with $\mathrm{H_{A:NI}}$; e.g., impute μ to missing values for R and $(\mu - \Delta_{NI})$ to missing values for T where μ can be an optimistic, median, or pessimistic value; for a dichotomous outcome, $\mu = 1$ is of interest where 1 corresponds to favorable outcome)
3. Homogeneity of treatment differences across subgroups
4. Management of interim analyses
5. Parallel or crossover designs
6. Centers as random source of variation

A major concern for non-inferiority clinical trials is "biocreep;" i.e., the tendency for recently demonstrated non-inferior treatments to be the active reference control treatments in new clinical trials even though they are actually somewhat inferior to historically proven active reference control treatments relative to placebo.
A hypothetical example is as follows:

1. About 20 years ago, R0 was proven to be significantly superior to placebo P in a clinical trial with success rates of 0.96 for 200 patients with R0 and 0.50 for 100 patients with placebo; the lower limit of two-sided 0.95 confidence interval for (R0 - P) is 0.36.
2. About 15 years ago, R1 was proven to be non-inferior to R0 in a clinical trial for which the two-sided 0.95 confidence interval indicated that the success rate of 0.91 for 600 patients with R1 was not worse than that of 0.96 for 300 patients with R0 by more than 0.10 (which is less than 30% of (R0 - P)).
3. About 10 years ago, R2 was proven to be non-inferior to R1 in a clinical trial for which a two-sided confidence interval indicated that the success rate of 0.86 for 800 patients with R2 was not worse than 0.91 for 400 patients with R1 by more than 0.10.
4. About 5 years ago, R3 was proven to be non-inferior to R2 in a clinical trial for which a two-sided 0.95 confidence interval indicated that the success rate of 0.81 for 1000 patients with R3 was not worse than 0.86 for 500 patients with R2 by more than 0.10.
5. Today, a clinical trial is being planned to demonstrate that R4 is non-inferior to R3 by being no more than 0.10 worse by a two-sided 0.95 confidence interval. From a meta-analysis for (1) - (4), the lower limit of the two-sided 0.95 confidence interval for (R3 - P) is 0.19 for which 0.10 is more than 50%. From a meta-analysis for (2) - (4), R3 is significantly inferior to R0 (two-sided p<0.05). Is this trial justifiable?

Major Issues for Confirmatory Clinical Trials to Demonstrate Efficacy Through Non-Inferiority (Department of Health and Human Services, FDA [1999], D'Agostino Sr., Massaro, and Sullivan [2003])

1. Clarity of well evident efficacy for at least one treatment
   a. superiority of T and/or R to placebo P (in $T,\ R,\ P$ study)
   b. superiority of a higher dose of T to a lower dose (in T1, T2, R study)
   c. superiority of R to historical experience for placebo
      i. historical comparison to placebo for experience with R
      ii. patient population is comparable to those for previous studies of R
      iii. data quality and study compliance are comparable to prior studies of R
2. Extent of potential inferiority that does not require preservation (i.e., non-inferiority boundary or margin)
   a. a generally agreed amount (e.g., differences in bioavailability parameters $\leq 20\%$ for bioequivalence, differences in healing rates $\leq 10\%$ in anti-infective studies)
   b. a fraction of the expected difference between R and P; i.e., $(1 - L)(R - P)$ with L being the fraction of $(R - P)$ which must have preservation in the demonstration of non-inferiority
3. Demonstrating superiority subsequent to non-inferiority is possible

The guidance in ICHE10 (Choice of control group in clinical trials) emphasizes assay sensitivity (i.e., "the ability to distinguish an effective treatment from a less effective treatment or an ineffective treatment").

1. Assay sensitivity applies directly for a valid demonstration of superiority of test treatment to a control through the observed significant difference between treatments.
2. A clinical trial for the demonstration of non-inferiority of a test treatment to an active reference control has its assay sensitivity strengthened when its conduct has sufficiently high quality and its structure is as similar as possible to the historical trials that demonstrated efficacy for the active reference control as follows:
   a. Patient population
   b. The actual form (or regimen) of the active reference control
   c. The assessments made for patients
3. ICHE10 indicates that assay sensitivity can be undermined by
   a. "Poor compliance with therapy"
   b. "Poor responsiveness of the enrolled study population to drug effects"
   c. "Use of concomitant non-protocol medication or other treatment that interferes with the test drug or that reduces the extent of the potential response"
   d. "An enrolled population that tends to improve spontaneously, leaving no room for further drug-induced improvement"
   e. "Poorly applied diagnostic criteria (patients lacking the disease to be studied)"
   f. "Biased assessment of endpoint because of knowledge that all patients are receiving a potentially active drug"
4. When assay sensitivity is arguably applicable, demonstration that the inferiority of test treatment relative to active reference control does not exceed a clinically appropriate bound (e.g., $\leq$ 50% (active reference control - placebo)) also demonstrates the hypothetical superiority of the test treatment relative to placebo and thereby its efficacy.

Statistical Methods to Demonstrate Superiority, Non-inferiority or Equivalence

1. Two-sided confidence interval
   a. For superiority when big is better, the lower inferential bound needs to exceed 0; the upper bound is descriptive.
   b. For non-inferiority when big is better, the lower inferential bound needs to exceed $-\Delta_{NI}$; the upper bound is descriptive.
   c. For equivalence, the entire confidence interval needs to be internal to $(-\Delta_E, \Delta_E)$; both the lower bound and the upper bound are inferential.
2. One (or two) tests of the one-sided null hypotheses that correspond to no superiority, inferiority, or no equivalence
3. Considerations for confidence intervals
   a. straightforward to construct and to interpret for comparisons between means and proportions
   b. straightforward to construct for odds ratios and hazard ratios, but can be awkward to interpret
   c. can be difficult to construct and to interpret for comparisons based on rankings of responses

Study to compare finasteride (R) and a plant extract (T) in 1098 men with benign prostate hyperplasia (Carraro et al [1996], Koch, Davis, and Anderson [1998])

1. Activity for R demonstrated by $p < 0.01$ for reduction of prostate volume, reduction of serum prostate antigen (PSA) levels, poorer sexual function
2. For Int. Prostate Symptom Score (IPSS), the confidence interval for $(T - R)$ was $(-0.17, 0.96)$ for change from baseline to week 26 which is internal to $(-2, 2)$
3. Non-inferiority of T to R is reasonably well supported (given no superiority comparison to placebo) since $SD = 6$ and $(\Delta\ /SD) = (2/6) = 0.33$

In comparisons of standard therapy to standard therapy plus new agent (or a new test agent to a standard agent) in areas such as

1. oncology
2. organ transplantation
3. cardiovascular disease

there can be uncertainty as to how much better the standard therapy is to placebo because data on efficacy relative to placebo is not available or is no longer relevant. This can undermine the demonstration of efficacy by a non-inferiority study and thereby make a superiority study necessary.

Considerations for the use of the three treatment design with T, R, and P [Koch A, and Rohmel J. Journal of Biopharmaceutical Statistics, 2004].

1. When assumptions regarding the historical trials for the reference treatment and placebo cannot be met, then a 3-armed "gold standard" trial is necessary to compare T, R, and P.
2. Reasons for including Placebo treatment (P) in the clinical trial
   a. R is a "traditional" standard, but there are doubts for its current efficacy, perhaps because trials were conducted too long ago to be applicable now
   b. R is a "weak" standard - (i.e., efficacy over placebo could be small)
   c. R is a "volatile" standard - (i.e., historical trials have produced widely varying placebo versus standard differences)
   d. Disease is not fully understood
3. Reasons for including an (R)
   a. Current reference treatment (R) might outperform the test treatment (T)
   b. If efficacy cannot be established with T compared to P, it is useful to know if R also failed compared to P so as to understand trial assay sensitivity (or validity).
4. T can be deemed successful in this "gold standard" trial if T is shown to be superior to P and noninferior to R, regardless of any other testing.
5. The two tests in (4) can be done without any adjustment to the significance level

Three treatment groups: test (T), active reference (R) control, placebo (P); see Koch and Tangen [1999].

1. Assess test vs placebo at one sided 0.025 significance level first
2. If (1) significant, put 0.95 lower confidence interval on $(T - P)/(R - P)$ by method for ratio estimator (e.g., Fieller); if lower bound L exceeds 0.50 (or 0.60), interpret T as "meaningfully better" than P; if L exceeds 0.67 (or 0.75), interpret T as "at least as good" as R; if L exceeds one, interpret as "weakly superior"; if L exceeds 1.5 (or 1.33), interpret as "superior."
3. If (2) supports "at least as good as" or more, evaluate reference vs placebo at one-sided 0.025 significance level
4. Consider 2 : 2 : 1 or 3 : 2 : 1 sample size allocation to test, reference, placebo since $T - P, R - P >$ (non-inferiority margin for $(T - R)$).

Study to compare T, R, and P for healing duodenal ulcer (hypothetical)

1. 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 is at least as good as R
5. Evidence is comparable to two studies

Planned integrated analysis for two multi-center studies to compare three treatments for rates of an unfavorable gastrointestinal outcome

1. Two multi-center studies with three randomly assigned treatments. The treatments were placebo (P), reference control (R), and test drug (T)
2. The primary objective with highest priority for each study was to show that T had lower rates of an unfavorable gastrointestinal outcome than R and that T provided better pain relief than P.
3. Given (2), the next objective was to show that T was non-inferior to P for rates of the unfavorable outcome for the combined studies. Low event rates for P and possibly T led to the combined studies being its planned basis. A criterion for non-inferiority of T to P was $(R - T)/(R - P) \geq 0.75$ from a one-sided lower 0.975 confidence interval. i.e., T preserved at least 75% of the reduction in unfavorable outcomes for R that P provided.

Sample sizes (n_T, n_R, n_P) for T, R, and P (and $(n_T + n_R + n_P) = n$ in total) in a clinical trial with $(1-\beta)$ power to demonstrate that $(T-P)/(R-P) > L$ significantly applies with one-sided $p < \alpha$ when $(T-P)/(R-P) = \theta > L$.

$$n_T = \frac{(Z_\alpha+Z_\beta)^2\left\{1+\frac{L^2}{c_R}+\frac{(1-L)^2}{c_P}\right\}\sigma^2}{(\theta-L)^2\Delta_{RP}^2}$$

where $n_R = c_R n_T, n_P = c_P n_T$; Z_α and Z_β are $100(1-\alpha)$ and $100(1-\beta)$ percentiles of the standard normal distribution with mean 0 and variance 1; σ^2 is the applicable variance; and Δ_{RP} is the expected difference between R and P. With $\alpha = 0.025$ and $\beta = 0.100$ for 0.90 power at one-sided 0.025 significance level, a clinical trial with $c_R = 1, c_P = 0.5$ (or 2 : 2 : 1 allocation) can contradict $L < 0.667$ when $\theta = 1.167$ and $(\Delta_{RP}/\sigma) = 0.600$ with

$$n_T = \frac{(1.96+1.282)^2(1+(0.667)^2+2(0.333)^2)}{(0.500)^2(0.600)^2} = 194 = n_R = 2n_P,\ n = 485$$

Sample size to demonstrate superiority, non-inferiority, or equivalence from the comparison of means (or proportions) for test drug (T) and control drug (R)

$$n \text{ per group} = \frac{(Z_\alpha+Z_\beta)^2(\sigma_T^2+\sigma_R^2)}{(\delta-\Delta)^2}.$$

Here σ_T^2 and σ_R^2 are the applicable variances for T and R; $\delta = (T-R)$ is the true difference between T and R; $Z_\alpha$and Z_β are the $100(1-\alpha)$ and $100(1-\beta)$ percentiles of the standard normal distribution, and they correspond to one-sided significance level α and one-sided power $(1-\beta)$. For $\alpha = 0.025$, $Z_\alpha = 1.96$; and for $(1-\beta) = 0.90$, $Z_\beta = 1.28$.

1. Superiority: $\Delta = 0$
2. Non-inferiority: $\Delta = -\Delta_{NI}$
3. Equivalence: $\Delta_1 = -\Delta_E$ and $\Delta_2 = \Delta_E$ and power is at least $(1-2\beta)$ when n is the maximum of the values corresponding to Δ_1 and Δ_2

The δ in a study to demonstrate superiority to placebo is usually two to three times as large as Δ_{NI} or Δ_E, and the $\mid \delta \mid$ in a study to demonstrate non-inferiority or equivalence is usually smaller than Δ_{NI} and Δ_E. Thus, sample sizes in studies to demonstrate non-inferiority or equivalence to an active control are usually much larger than those to demonstrate superiority to placebo.

For study to compare T and R to show $(T-R) \geq -0.333\Delta_{RP} = -0.200\sigma$ when $(T-R) = 1.167(R-P) - (R-P) = 0.167(R-P) = 0.167(0.60\sigma) = 0.10\sigma$

$$n_T = \frac{(1.96+1.282)^2(2\sigma^2)}{(0.30)^2\sigma^2} = 233 = n_R,\ n = 466$$

Purposes served by analysis of covariance

1. More powerful statistical test (or narrower confidence interval) through "variance reduction" in statistic for comparison of randomized groups
2. Conduct of comparison between randomized groups in setting for which random imbalances for covariables are "adjusted to equivalence"
3. Clarify the degree to which detected differences between randomized groups are due to treatment rather than other factors which are associated with response
4. Provides some structure for evaluating homogeneity of treatment differences across subgroups

Covariables for adjustment

1. a priori specification is necessary to avoid spurious role
2. strong correlation with response criteria provides variance reduction and increased power
3. non-parametric methods have minimal assumptions

Methods for covariance analysis

1. Parametric through statistical models for the relationship between covariables and the conditional distributions of response given the covariables
2. Nonparametric through linear models for (unconditional) differences between treatment groups for response criteria and covariables jointly and with specifications that adjust differences for covariables to 0.
   a. for tests of no difference between treatment groups, the study design provides the basis for the distribution of results under the null hypothesis
   b. for confidence intervals concerning treatment differences and tests concerning treatment × subgroup interaction, both randomized assignment and patient selection being comparable to a simple random sample for each treatment × subgroup are needed.

Dental Clinical Trial To Compare Three Treatments For Reducing Dental Plaque Scores (Hadgu, A. and Koch, G.G. [1999, Journal of Biopharmaceutical Statistics])

1. The trial included 109 patients with preexisting plaque, but without periodontal disease, and a minimum of 20 sound natural teeth
2. Patients were randomly assigned in a double masked way to use of a control (C), reference (R), or test (T) mouth rinse during each day of a 6 month follow-up period
3. The primary response variables were the plaque scores at 3 months, 6 months, and their average
4. The plaque score at baseline was an important covariable which had strong correlations with the response variables
5. Gender, age, and smoking status were background variables which had essentially no association with the response variables
6. Missing values at 6 months for 4 patients were replaced by values at 3 months

Means and Standard Errors (S.E.) of Plaque Scores at Baseline, 3 Months, and 6 Months for Patients in Dental Clinical Trial

Visit	Statistic	Control (C)	Reference (R)	Test (T)
		($n = 39$)	($n = 34$)	($n = 36$)
Baseline	Mean	2.562	2.569	2.479
	S.E.	0.055	0.061	0.049
3 Months	Mean	1.786	1.315	1.255
	S.E.	0.112	0.123	0.092
6 Months	Mean	1.763	1.243	1.032
	S.E.	0.096	0.127	0.075

Results From Unadjusted Treatment Comparisons For Dental Clinical Trial

Comparison	Statistic	3 Months	6 Months	Average
$T - P$	Estimate	0.530	0.731	0.631
	S.E.	0.145	0.122	0.116
	p-value	< 0.001	< 0.001	< 0.001
$R - P$	Estimate	0.470	0.520	0.495
	S.E.	0.166	0.159	0.149
	p-value	0.005	0.001	0.001
$\frac{T-P}{R-P}$	Estimate	1.128	1.406	1.274
	(Confidence	(0.603,	(0.893,	(0.819,
	Interval (0.95))	2.891)	3.081)	2.645)
	p-value (1.0)	0.695	0.152	0.300
For treatment × visit, $p = 0.314$ from d.f. $= 2$ test.				

Results From Covariance Adjusted Treatment Comparisons For Dental Clinical Trial

Comparison	Statistic	3 Months	6 Months	Average
$T - P$	Estimate	0.449	0.684	0.567
	S.E.	0.118	0.109	0.094
	p-value	< 0.001	< 0.001	< 0.001
$R - P$	Estimate	0.454	0.528	0.491
	S.E.	0.139	0.131	0.119
	p-value	0.001	< 0.001	< 0.001
$\frac{T-P}{R-P}$	Estimate	0.989	1.296	1.154
	(Confidence	(0.511,	(0.847,	(0.760,
	Interval (0.95))	2.193)	2.342)	2.028)
	p-value (1.0)	0.971	0.241	0.516
For treatment × visit, $p = 0.190$ from d.f. $= 2$ test.				

Methodology for nonparametric analysis of covariance for a randomized clinical trial without stratification

Group	Sample Size	Mean Response	Means for m Covariables
Control (C)	n_C	$\bar{y}_C$	$\bar{\boldsymbol{x}}_C$
Reference (R)	n_R	$\bar{y}_R$	$\bar{\boldsymbol{x}}_R$
Test (T)	n_T	$\bar{y}_T$	$\bar{\boldsymbol{x}}_T$
$(T-C)$		$d_{TC} = \bar{y}_T - \bar{y}_C$	$(\bar{\boldsymbol{x}}_T - \bar{\boldsymbol{x}}_C) = \boldsymbol{u}_{TC}$
$(R-C)$		$d_{RC} = \bar{y}_R - \bar{y}_C$	$(\bar{\boldsymbol{x}}_R - \bar{\boldsymbol{x}}_C) = \boldsymbol{u}_{RC}$

Use weighted least squares to fit the linear model

$$E(\boldsymbol{F}) = E\begin{bmatrix} d_{TC} \\ \boldsymbol{u}_{TC} \\ d_{RC} \\ \boldsymbol{u}_{RC} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \mathbf{0} & \mathbf{0} \\ 0 & 1 \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \gamma_{TC} \\ \gamma_{RC} \end{bmatrix} = \boldsymbol{X\gamma}$$

The weights are from the estimated covariance matrix $\boldsymbol{V_F}$ for $\boldsymbol{F}$.

The covariance adjusted estimates for differences from control are

$$\boldsymbol{g} = \begin{bmatrix} \hat{\gamma}_{TC} \\ \hat{\gamma}_{RC} \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{F}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}_{\boldsymbol{F}}^{-1}\boldsymbol{F}$$

and Their Estimated Covariance Matrix is $\boldsymbol{V_g} = (\boldsymbol{X}'V_F^{-1}\boldsymbol{X})^{-1}$, where

$$\boldsymbol{V_F} = \begin{bmatrix} \boldsymbol{V}_C + \boldsymbol{V}_T & \boldsymbol{V}_C \\ \boldsymbol{V}_C & \boldsymbol{V}_C + \boldsymbol{V}_R \end{bmatrix}, \text{where}$$

$$\boldsymbol{V}_i = \frac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i} \begin{bmatrix} (y_{ij} - \bar{y}_i)^2 & (y_{ij} - \bar{y}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)' \\ (y_{ij} - \bar{y}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) & (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)' \end{bmatrix}$$

The adjusted estimates $\boldsymbol{g}$ have an approximately bivariate normal distribution.

For the hypothesis, $H_0 : \boldsymbol{C\gamma} = \mathbf{0}$ with $\boldsymbol{C}$ full rank

$Q(\boldsymbol{Cg}) = \boldsymbol{g}'\boldsymbol{C}'\{\boldsymbol{CV_gC}'\}^{-1}\boldsymbol{Cg}$

approximately has chi-squared distribution with d.f. $=$ Rank($\boldsymbol{C}$), where

$\boldsymbol{C} = [1, 0]$ for T vs C

$\boldsymbol{C} = [0, 1]$ for R vs C

$\boldsymbol{C} = [1, -K]$ for $(T - C)/(R - C) = K$

With $0.333 \leq K \leq 3.00$, the similarity of T and R is evaluated. Fieller's formula yields a confidence interval for $(T - C)/(R - C)$. The extent to which the model counteracts random imbalance is evaluated with $Q = (\boldsymbol{F} - \hat{\boldsymbol{F}})'V_F^{-1}(\boldsymbol{F} - \hat{\boldsymbol{F}})$ where $\hat{\boldsymbol{F}} = \boldsymbol{Xg}$. It is a goodness of fit statistic which approximately has the chi-squared distribution with d.f. $= 2m$. Extensions to account for stratification in randomized assignments are available.

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 α 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 α.
   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 ≥ 3 or the number of hypotheses for the second objective is ≥ 2

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 $(H_{0H1}\bigcap H_{0L1})\bigcup H_{0\mathrm{L1}}\bigcup H_{0\mathrm{L2}} = H_{0\mathrm{L2}}$
   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}, $H_{0H1}\bigcap H_{0L1}$, $H_{0H1}\bigcap H_{0\mathrm{H2}}$=$H_{0H1}$, $H_{0H1}\bigcap H_{0\mathrm{L2}}$; 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}

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.

Berger, R.L. and Hsu, J.C. [1996]. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, 11:283-319.

Blackwelder, W.C. [1982]. Proving the null hypothesis in clinical trials. Controlled Clinical Trials, 3:345-353.

Carraro, J.C., Raynaud, J.P., Koch, G.G., et al. [1996]. Comparison of phytotherapy (permixon) with finasteride in the treatment of benign prostate hyperplasia: a randomized international study of 1,098 patients. The Prostate, 29:231-240.

Chen, G., Wang, Y.-C., Chi, G. YH. [2004]. Hypotheses and type I error in active-control noninferiority trials. Journal of Biopharmaceutical Statistics, 14(2): 301-314.

Cheuvart, B., Bollaerts, A. [1999]. Sample size considerations for assessing vaccine consistency through equivalence. Drug Information Journal, 33:149-152.

Chuang-Stein, C. [1999]. Clinical equivalence-a clarification. Drug Information Journal, 33:1189-1194.

Chuang-Stein, C., Sanders, C., and Snapinn, S. [2004]. An industry survey on current practices in the design and analysis of active control studies. Journal of Biopharmaceutical Statistics, 14(2): 349-358.

D'Agostino Sr., R.B., Massaro, J.M., and Sullivan, L.M. [2003]. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. Statistics in Medicine, 22:169-186.

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.

Department of Health and Human Services, FDA. [1999]. International Conference on Harmonisation; E10: Choice of Control in Clinical Trials. Federal Register Vol. 64; No. 185.

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.

Durrleman, S. and Simon, R. [1990]. Planning and monitoring of equivalence studies. Biometrics, 46: 329-336.

Fisher L.D, Gent, M., Büller H.R. [2001]. Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin and placebo, American Heart Journal, 141: 26-32.

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

Gillings, D. and Koch, G.G. [1991]. An application of the principle of intent-to-treat to the analysis of clinical trials, Drug Information Journal, 25, 411-424.

Hasselblad, V. and Kong, D.F. [2001]. Statistical methods for comparison to placebo in active-control trials, Drug Information Journal, 35: 435-449.

Hauck, W.W. and Anderson, S. [1999]. Some issues in the design and analysis of equivalence trials, Drug Information Journal, 33:109-118.

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

Holmgren, E.B. [1999]. Establishing equivalence by showing that a specified percentage of the effect of the active control over placebo is maintained. Journal of Biopharmaceutical Statistics, 9(4): 651-659.

Hung, H.M.J. and Wang, S.-J. [2004]. Multiple testing of noninferiority hypotheses in active controlled trials. Journal of Biopharmaceutical Statistics, 14(2): 327-336.

Hung, H.M.J., Wang, S.-J., Tsong, Y., Lawrence, J., and O'Neill, R.T. [2003]. Some fundamental issues with non-inferiority testing in active controlled trials. Statistics in Medicine, 22:213-226.

Hwang, I.K. and Morikawa, T. [1999]. Design issues in noninferiority/equivalence trials. Drug Information Journal, 33: 1205-1218.

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

Jones, B., Jarvis, P., Lewis, J.A., Ebbutt, A.F. [1996]. Trials to assess equivalence: the importance of rigorous methods, British Medical Journal, 313: 36-39.

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. [2002]. Missing data: a damaging (and possibly catastrophic disorder) for statistical planning. Bulletin of the International Chinese Statistical Association, 43-45.

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

Kong, L., Kohberger, R.C., and Koch, G.G. [2004]. Type I error and power in non-inferiority/equivalence trials with correlated multiple endpoints: an example from vaccine development trials. Journal of Biopharmaceutical Statistics, In press.

Laster, L.L. and Johnson, M.F. [2003]. Non-inferiority trials: the 'at least as good as' criterion. Statistics in Medicine, 22(2): 187-200.

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

Lynch, C.J. and Lachenbruch, P.A. [1996]. Statistical issues in biologics submissions to the FDA. Drug Information Journal, 30:921-932.

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.

Rothmann, M., Li, N., Chen, G., Chi, G.Y.H., Temple, R., and Tsou, H.-H. [2003]. Design and analysis of non-inferiority mortality trials in oncology. Statistics in Medicine, 22:239-264.

Schell, M.J., McBride, M.A., Gennings, C., and Koch, G.G. [2001]. The intention-to-treat principle for clinical trials. Chapter 12 in Clinical Trials in Neurology, ed. R.J. Guiloff. Springer-Verlag, London, pp. 131-144.

Schuirmann, D. [1990]. Design of bioavailablility/bioequivalence studies. Drug Information Journal, 24: 213-224.

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

Siegel, J.P. [2000]. Equivalence and non-inferiority trials. American Heart Journal, 139: s166-s170.

Simon, R. [1999]. Bayesian design and analysis of active control clinical trials. Biometrics, 55: 484-487.

Snapinn, S.M. [2004]. Alternatives for discounting in the analysis of noninferiority trials. Journal of Biopharmaceutical Statistics, 14(2): 263-274.

Tang, M.-L. and N.-S. Tang [2004]. Tests of noninferiority via rate difference for three-arm clinical trials with placebo. Journal of Biopharmaceutical Statistics, 14(2): 337-348.

Temple, R., Ellenberg, S.S. [2000]. Placebo-controlled trials and active-control trials in the evaluation of new treatments - Part 1: Ethical and scientific issues, Part 2: Practical issues and specific cases. Annals of Internal Medicine, 133: 455-463, 464-470.

Tsong, Y., Wang, S.J., Hung, H.M.J., and Cui, L. [2001]. Objectives, designs and analysis of non-inferiority active controlled clinical trials. Proceedings of Biopharmaceutical Section of American Statistical Association.

Wang, S.-J. and Hung, H.M.J. [2003]. TACT method for non-inferiority testing in active controlled trials. Statistics in Medicine, 22:227-238.

Wang, S.J., Hung, H.M.J., and Tsong, Y. [2002]. Utility and pitfall of some statistical methods in active controlled clinical trials. Controlled Clinical Trials, 23: 15-28.

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.