ON WILCOXON COMPARISONS FOR MULTISITE CLINICAL TRIALS

Dennis D. Boos and Cavell Brownie

Institute of Statistics Mimeo Series No. 1945

April, 1989

On Wilcoxon Comparisons for Multisite Clinical Trials

Dennis D. Boos and Cavell Brownie

Department of Statistics

North Carolina State University

Raleigh, NC 27695 - 8203

SUMMARY

New rank-based methods for analyzing data from multisite clinical trials are presented in the context of "mixed" linear models. In contrast to current rank methods, the new procedures test for a drug main effect in the presence of a random drug by site interaction (or *drug by investigator* interaction when there is only one investigator per site). Analogous procedures are also provided for the "fixed effects" situation, and comparisons are made with current methods. The rationale for an analysis which assumes random investigator effects is described.

Key words: Drug by investigator interaction; Homogeneity test; Mann-Whitney U; Mantel-Haenszel test; Mixed models; Ordered categorical data; Rank tests; van Elteren test.

1. Introduction

Clinical trials to compare a new drug with a standard often involve the participation of a number of investigators at different sites. For each investigator (or site), subjects are assigned randomly to the new drug or standard, so that the experimental design can be viewed as a randomized block design with each treatment (drug) replicated several times in each block. For simplicity we assume that there is only one investigator at each site so that a block may be described as either an investigator or a site.

In assessing performance of a new drug, questions of interest relate to whether there is an overall difference between drugs across investigators (a drug main effect) and whether the difference is reasonably consistent (no substantial drug by investigator interaction). Using "linear model" terminology and standard practice, investigator effects should be regarded as "fixed" only if inferences about drug performance are to be limited to just those investigators in the study. The more likely scenario is that it is of interest to make inferences about drug performance for a larger population of clinicians. In that case the investigator effects would be viewed as random even though selection of participating investigators is not at random. This is because inferences based on investigators as the relevant level of replication are more likely to reflect potential performance in a broader population than are procedures which ignore the components of variation due to clinicians. To test for a drug main effect without ruling out a drug by investigator interaction, it is thus necessary to specify whether the first-mentioned "fixed effects" model or the latter "mixed" model is appropriate.

In this article we consider rank-based alternatives to the analysis of variance (ANOVA) F tests for analyzing data from multisite clinical trials, with emphasis on the mixed model situation. Our work was prompted by the realization that currently popular methods such as van Elteren's test (see, e.g., Lehmann, 1975, p. 145) are not appropriate if there is a random drug by investigator interaction. To assume the absence of such interaction seems

unreasonable because the presence of interaction is often itself a question of major importance. Our objective is therefore to introduce a simple rank-based procedure for the problem of testing for a drug main effect in the presence of random investigator and possibly drug by investigator effects. This new procedure is developed in Section 2 and compared to currently used methods. In Section 3 we consider the fixed effects situation and propose new procedures for detecting main and interaction effects. Problems with the interpretation of current methods are also noted. Section 4 contains a real data example that illustrates the major difference between our procedure and the van Elteren test, and Section 5 provides additional motivation for the mixed model analysis. An Appendix contains details of variance estimates.

2. Analysis for the Mixed Model Situation

2.1 Data Types

Although ANOVA F tests are remarkably robust to nonnormality with respect to test size, the use of rank-based procedures can often be justified either because they result in increased power for outlier-prone (or heavy-tailed) data or because responses are measured on an ordinal rather than a continuous scale. Examples of ordinal data, more specifically ordered categorical data, are especially common in clinical trials where patient response is often recorded as one of several ordered outcomes such as "condition deteriorated," "no change," ..., "marked improvement." We are therefore particularly interested in rank-based procedures for ordered categorical data, but have tried to make the development as general as possible, allowing for responses that are continuous or ordered categorical.

2.2 Model Formulation

For the i*th* investigator, i=1, ..., a, we let X_{ij}, j=1, ..., m_i be the response for the j*th* patient on drug 1 and Y_{ik}, k=1, ..., n_i be the response for the k*th* patient on drug 2. The usual linear

model with random block and interaction effects is

$$X_{ij}=\mu+\alpha_i+\beta_1+\gamma_{i1}+\epsilon_{xij},$$

$$Y_{ik}=\mu+\alpha_i+\beta_2+\gamma_{i2}+\epsilon_{yik}, \tag{2.1}$$

where μ is an unknown constant, β_1 and β_2 ($\beta_2=-\beta_1$) are fixed drug effects, and α_i, γ_{ij} are random effects for investigator and drug by investigator interaction, respectively. We assume that the α_i are iid $(0, \sigma_\alpha^2)$, the γ_{ij} are iid $(0, \sigma_\gamma^2)$, the ϵ_{xij} and ϵ_{yik} are iid $(0, \sigma_\epsilon^2)$, and that all the random variables are mutually independent.

At first glance, the notation in (2.1) may seem awkward since the more standard linear model notation is to write Y_{i1k} and Y_{i2k} for responses to drugs 1 and 2 respectively (with n_{i1} and n_{i2} the respective sample sizes). Our use of X_{ij} and Y_{ik} for Y_{i1j} and Y_{i2k} (and of m_i, n_i for n_{i1}, n_{i2}) is to achieve as much consistency as possible with notation that is common in the literature on two sample rank procedures. This should improve readability, especially for those who are used to the X, Y and m, n notation.

A measure of the difference between drugs for the ith investigator based on average response is

$$\Delta_i=E(Y_{i1}-X_{i1}).$$

Note that for the model (2.1) $\Delta_i=\beta_2-\beta_1=2\beta_2$, i=1, ..., a. A second measure of this effect is

$$\theta_i=P(Y_{i1}>X_{i1}) \quad \text{for continuous X and Y,} \quad \text{or}$$

$$\theta_i=P\Big(g(Y_{i1})>g(X_{i1})\Big)+\tfrac{1}{2}\,P\Big(g(Y_{i1})=g(X_{i1})\Big),$$

where g(X) and g(Y) represent the ordered categorical responses obtained by discretizing the

often unobservable X and Y. The measure θ is particularly appealing in clinical trials because θ can be thought of as the probability of a more favorable response with drug 2 than with drug 1 for a randomly chosen patient.

Tests of the null hypothesis of no drug main effect, H_{01}: $\beta_1=\beta_2=0$, may be constructed using estimators of Δ_i as in ANOVA or using Wilcoxon-Mann-Whitney (WMW) type estimators of θ_i. For unbalanced data (and assuming normality) ANOVA does not in general lead to an exact test, but there is a large literature on approximate tests (see, e.g., Webster, 1968). The test we introduce is based on estimators of θ_i and is analogous to an approximate F test arising from Yates' (1934) unweighted means analysis (e.g., Webster, 1968).

2.3 The New Test for a Drug Main Effect

For the ith investigator the WMW estimator of $\hat{\theta}_i$ is

$$\begin{aligned}\hat{\theta}_i &= \{(\#\text{ of pairs } (X_{ij}, Y_{ik}) \text{ with } Y_{ik}>X_{ij}) \\ &\quad +\tfrac{1}{2}(\#\text{ of pairs } (X_{ij}, Y_{ik}) \text{ with } Y_{ik}=X_{ij})\}/(m_i n_i) \\ &= \frac{1}{m_i n_i} \sum_{j=1}^{m_i} \sum_{k=1}^{n_i} [\, I\,(Y_{ik}>X_{ij})+\tfrac{1}{2}\, I(Y_{ik}=X_{ij})].\end{aligned}$$

If ordered categorical data are observed, $\hat{\theta}_i$ is defined with $g(X_{ij})$ and $g(Y_{ik})$ in place of X_{ij} and Y_{ik}. In either case $\hat{\theta}_i$ is just $[W_i^*-n_i(n_i+1)/2]/m_i n_i$, where W_i^* is the Wilcoxon rank sum for the ith investigator when mid-ranks are used to break ties (see Lehmann, 1975, p. 22). Clearly $E(\hat{\theta}_i)=\theta_i$ due to the U-statistic structure of $\hat{\theta}_i$. Also, note that $\beta_1=\beta_2=0$ implies that $\theta_1=\theta_2=\cdots=\theta_a=\frac{1}{2}$, due to the fact that each pair (X_{ij}, Y_{ik}) is then exchangeable. Thus for the mixed model (2.1), the null hypothesis of no treatment main effect, H_{01}: $\beta_1=\beta_2=0$, leads

to a simple and interpretable hypothesis in the θ_i. This is in vivid contrast to the fixed effects model discussed in Section 3.

To construct a test of $\theta_1=\theta_2=\cdots=\theta_a=\frac{1}{2}$, it is important to note that the variance of $\hat{\theta}_i$ is affected by the random drug by investigator effect which enters into the comparisons $I(Y_{ik} > X_{ij})$. For continuous data from the model (2.1), the variance of $\hat{\theta}_i$ is given by

$$\begin{aligned} \text{Var } \hat{\theta}_i &= [\theta_i(1-\theta_i)+(n_i-1)(p_{2i}-\theta_i^2)+(m_i-1)(p_{3i}-\theta_i^2) \\ &\quad +(m_i-1)(n_i-1)(p_{4i}-\theta_i^2)]/m_i n_i, \end{aligned} \tag{2.2}$$

where $\theta_i=P(Y_{i1}>X_{i1})$, $p_{2i}=P(Y_{i1}>X_{i1}, Y_{i2}>X_{i1})$, $p_{3i}=P(Y_{i1}>X_{i1}, Y_{i1}>X_{i2})$, and $p_{4i}=P(Y_{i1}>X_{i1}, Y_{i2}>X_{i2})$. (See, Hettmansperger, 1984, p. 158). If $\theta_i=\frac{1}{2}$, and $p_{2i}=p_{3i}=\frac{1}{3}$, then Var $\hat{\theta}_i$ reduces to

$$\left(\frac{m_i+n_i+1}{12m_i n_i}\right) + \frac{(m_i-1)(n_i-1)(p_{4i}-\frac{1}{4})}{m_i n_i}. \tag{2.3}$$

The first term of (2.3) is the usual null variance of $\hat{\theta}_i$, $\sigma_0^2(\hat{\theta}_i)=(m_i+n_i+1)/12m_i n_i$, and the second term is the contribution due to the random drug by investigator interaction. Thus, inferences and methods based on $\sigma_0^2(\hat{\theta}_i)$ should not be used under (2.1). For example, the motivation for van Elteren's statistic (see Lehmann, 1975, p.145, or van Elteren, 1960)

$$V=\sum_{i=1}^{a} \frac{1}{\sigma_0^2(\hat{\theta}_i)}(\hat{\theta}_i-\tfrac{1}{2})$$

as an average of $\hat{\theta}_i$ weighted inversely proportional to variances, no longer holds. More serious is that the related test procedure based on comparison of

$$V_E=V \Big/ \left[\sum_{i=1}^{a}\left(1/\sigma_0^2(\hat{\theta}_i)\right)\right]^{\frac{1}{2}}$$

to a standard normal is not valid.

Our proposed test for main effects is to compare

$$V_1 = \frac{\sqrt{a}(\bar{\hat{\theta}} - \frac{1}{2})}{\sqrt{\frac{1}{a-1}\Sigma(\hat{\theta}_i - \bar{\hat{\theta}})^2}}$$

to a t distribution with $a-1$ degrees of freedom. Here $\bar{\hat{\theta}} = \Sigma\hat{\theta}_i/a$, so that V_1 is just a one sample t statistic computed on the $\hat{\theta}_i$. By the independence of the $\hat{\theta}_i$ we have that

$$E\left[\frac{1}{a-1}\sum_{i=1}^{a}(\hat{\theta}_i - \bar{\hat{\theta}})^2\right] = \frac{1}{a}\sum_{i=1}^{a}\text{Var}\,\hat{\theta}_i = \text{Var}(\sqrt{a}\bar{\hat{\theta}})\ .$$

Thus, V_1 is properly studentized with regard to expectation even though the $\hat{\theta}_i$ are not identically distributed except when the sample size pairs (m_i, n_i) are the same for each investigator. As $a \to \infty$, V_1 will converge to a standard normal random variable under H_{01}, but when a is small we suggest using the t_{a-1} distribution to obtain critical values. This is similar to Webster's (1968) use of the $F_{1,a-1}$ distribution to obtain critical values for a statistic $(V_\Delta)^2$, where V_Δ is defined as V_1 but with $\hat{\theta}_i$ replaced by $\hat{\Delta}_i = \bar{Y}_i - \bar{X}_i$, the difference between the sample means for the *ith* investigator. Basing critical values for V_1 or V_Δ on the t_{a-1} distribution should be reasonable if the individual $\hat{\theta}_i$ or $\hat{\Delta}_i$ have similar variances (see Table 1 and Webster, 1968).

Compared to V_E and other weighted averages of the form $\Sigma c_i\hat{\theta}_i$, the numerator of V_1 is approximately optimal under our mixed model (2.1) since the variance of $\hat{\theta}_i$ is nearly constant over investigators even with unbalanced sample sizes. To see this, consider first the analogous property for the difference in means,

$$\text{Var}(\bar{Y}_i - \bar{X}_i) = \text{Var}(\gamma_{i2} - \gamma_{i1} + \bar{\epsilon}_{yi} - \bar{\epsilon}_{xi}) = 2\sigma_\gamma^2 + \sigma_\epsilon^2(\frac{1}{m_i} + \frac{1}{n_i}).$$

If m_i and n_i are not small, then $2\sigma_\gamma^2$ is the dominant term and the variances for different investigators are approximately the same. Similarly, the variance of $\hat{\theta}_i$ in (2.3) will be dominated by $p_{4i}-\frac{1}{4}$ when m_i and n_i are not small or when the ratio $r=\sigma_\gamma^2/\sigma_\epsilon^2$ is not small. For a numerical example, if all the random variables are normally distributed, then $p_{2i}=p_{3i}$ and p_{4i} are straightforward to calculate from the bivariate normal (see Gupta, 1963). Table 1 shows a variety of situations where the variances of $\hat{\theta}_i$ and $\hat{\Delta}_i$ exceed the asymptotic values $p_{4i}-\frac{1}{4}$ and $2\sigma_\gamma^2$, respectively, by less than 25%. Thus we expect the test based on V_1 to be nearly optimal among tests based on averages of the $\hat{\theta}_i$ since the variances of the $\hat{\theta}_i$ are nearly equal.

2.4 The Test for Interaction.

For continuous data from (2.1) the null hypothesis of no interaction H_{02}: $\sigma_\gamma^2=0$ implies $\theta_1=\cdots=\theta_a=P(\epsilon_{y11}>\epsilon_{x11}-2\beta_2)$, where the common value θ_1 is different from $\frac{1}{2}$ unless $\beta_2=0$. To test H_{02} we suggest comparing

$$V_2=\sum_{i=1}^{a} w_i\,(\hat{\theta}_i-\hat{\theta}_w)^2$$

to the χ^2_{a-1} distribution, where w_i^{-1} is a consistent estimator of $\sigma^2_{02}(\hat{\theta}_i)$, the variance of $\hat{\theta}_i$ under H_{02}, and $\hat{\theta}_w=\Sigma w_i\hat{\theta}_i/\Sigma w_i$. To obtain the weights w_i, note that $\sigma^2_{02}(\hat{\theta}_i)$ is derived from (2.2) by setting $p_{4i}-\theta_i^2=0$ (e.g., Lehmann, 1975, p. 335-336) and also that under H_{02}, $p_{21}=\cdots=p_{2a}$ and $p_{31}=\cdots=p_{3a}$. This suggests pooling information across investigators (especially if m_i, n_i are small) to estimate the common values θ_1, p_2, p_3. Thus let

$$\hat{\theta}_d=\Sigma d_i\hat{\theta}_i,\ \hat{p}_{2e}=\Sigma\, e_i\hat{p}_{2i},\ \text{and}\ \hat{p}_{3f}=\Sigma f_i\hat{p}_{3i},$$

where $\Sigma d_i=\Sigma e_i=\Sigma f_i=1$, and for the i*th* investigator, $\hat{p}_{2i}$ and $\hat{p}_{3i}$ are U-statistic estimators of p_{2i} and p_{3i}, respectively [see equations (A3) in the Appendix]. For simplicity, we recommend

the weights

$$d_i=e_i=f_i=(m_in_i)/[(m_i+n_i+1)\Sigma(m_in_i)/(m_i+n_i+1)].$$

Then $w_i^{-1}=\hat{\sigma}^2_{02}(\hat{\theta}_i)$ is calculated from (2.2) with $p_{4i}-\theta_i^2=0$ and substituting estimates $\hat{\theta}_d$, $\hat{p}_{2d}$ and $\hat{p}_{3d}$ for the corresponding parameters. There are of course other candidates for estimating $\sigma^2_{02}(\hat{\theta}_i)$ which may have better properties than $\hat{\sigma}^2_{02}(\hat{\theta}_i)$ in some situations, but a complete discussion of this issue is beyond the scope of the present article.

For ordered categorical data, the expression for Var $\hat{\theta}_i$ under H_{02} is more complex [see Appendix equations (A2) and (A4)], but the same approach can be used to estimate this variance (see the Appendix for details). Finally, we note that the test on means for quantitative interaction proposed by Gail and Simon (1985) can be generalized to the $\hat{\theta}_i$ in an obvious fashion.

3. The Fixed Effects Model

3.1 Tests for Main and Interaction Effects

Here we discuss the situation where investigator main and interaction effects are assumed fixed as in the linear model

$$\begin{aligned} X_{ij}&=\mu+\alpha_i+\beta_1+\gamma_{i1}+\epsilon_{xij}, \\ Y_{ik}&=\mu+\alpha_i+\beta_2+\gamma_{i2}+\epsilon_{yik}, \end{aligned} \tag{3.1}$$

where $\Sigma_{i=1}^{a}\alpha_i=0$, $\beta_2=-\beta_1$, $\Sigma_{i=1}^{a}\gamma_{ij}=\Sigma_{j=1}^{2}\gamma_{ij}=0$, and the ϵ_{xij}, ϵ_{yik} are iid $(0, \sigma_\epsilon^2)$. Again it is important to test for a drug main effect without assuming interaction is absent. A test for interaction in the presence of a drug main effect is also of interest. For effects

measured in terms of mean response (i.e., in terms of the $\Delta_i=2\beta_2+2\gamma_{i2}$), the respective null hypotheses from (3.1) are

$$H_{01}:\ \beta_1=\beta_2=0 \text{ or equivalently } H_{01}:\ \tfrac{1}{a}\Sigma\Delta_i=\bar{\Delta}=0$$

and

$$H_{02}:\ \gamma_{i1}=\gamma_{i2}=0,\ i=1,\cdots,a \text{ or } H_{02}:\ \Delta_1=\Delta_2=\cdots=\Delta_a.$$

If ANOVA is judged appropriate, H_{01} and H_{02} are tested in the usual way using partial or SAS Type III sums of squares (e.g., Searle, 1987, p. 463) if m_i, n_i are not the same for each i.

Turning to effects measured in terms of θ_i, we see that in contrast to the mixed model (2.1), H_{01} for the fixed effects situation does not translate into a simple hypothesis in terms of the θ_i. For continuous data from (3.1),

$$\begin{aligned}\theta_i &= P(\mu+\alpha_i+\beta_2+\gamma_{i2}+\epsilon_{yik}>\mu+\alpha_i+\beta_1+\gamma_{i1}+\epsilon_{xij})\\ &= P(\epsilon_{yik}-\epsilon_{xij}>-2\beta_2-2\gamma_{i2})\\ &\equiv\theta(\beta_2,\gamma_{i2}).\end{aligned}$$

The equivalent constraints $\beta_1=\beta_2=0$ or $\bar{\Delta}=0$ under the parameterization of (3.1) are not reflected in a simple relationship among the θ_i because $\sum_{i=1}^{a}\gamma_{i2}=0$ does not lead to a simplification of $\sum_{i=1}^{a}\theta_i=\Sigma\theta(0,\gamma_{i2})$. Thus there is no general way to formulate H_{01}: $\bar{\Delta}=0$ as a hypothesis in terms of the θ_i, nor is it possible to derive test procedures for H_{01} based on the $\hat{\theta}_i$. In other words, given the model (3.1) a test based on the $\hat{\theta}_i$ cannot in general be interpreted as a test for a drug main effect if this is measured in terms of mean response.

We now suggest another way to view this problem motivated by the fact noted earlier that the θ_i have a particularly meaningful interpretation in clinical trials. For fixed investigators, i.e., inferences are limited to investigators involved in the study, it seems reasonable to interpret $\bar{\theta}=\frac{1}{a}\Sigma\theta_i>\frac{1}{2}$ as indicating a drug main effect. For c_i that represent

relative investigator patient population sizes, a null hypothesis of interest might be $\Sigma c_i(\theta_i - \frac{1}{2})=0$ for either continuous data or ordered categorical data. Usually, however, we would recommend $c_i=1/a$ as being preferable to the sample size dependent $c_i=m_i n_i/(m_i+n_i+1)$ in V_E. We therefore consider testing

$$H_{03}: \ \overline{\theta}=\tfrac{1}{2} \quad \text{or possibly} \quad H_{03}: \ \Sigma c_i(\theta_i-\tfrac{1}{2})=0$$

and

$$H_{04}: \ \theta_1=\theta_2=\cdots=\theta_a \ \text{ (no interaction).}$$

Test procedures based on the θ_i are now developed, the main difficulty being the problem of obtaining satisfactory estimators for Var $\hat{\theta}_i$ when the m_i and n_i are small.

For continuous data, the variance of $\hat{\theta}_i$ under (3.1) is

$$\sigma^2(\hat{\theta}_i)=\tfrac{1}{m_i n_i}[\theta_i(1-\theta_i)+(n_i-1)(p_{2i}-\theta_i^2)+(m_i-1)(p_{3i}-\theta_i^2)], \tag{3.2}$$

obtained from (2.2) by setting $p_{4i}-\theta_i^2=0$. For ordered categorical data, $\sigma^2(\hat{\theta}_i)$ is given in the Appendix [equations (A2) and (A4)] and also in Hochberg, (1981, p. 1726). Under H_{03}, the probabilities θ_i, p_{2i}, and p_{3i} are each dependent on i so that $\sigma^2(\hat{\theta}_i)$ must be estimated using data from the i*th* investigator only. Let $\hat{\sigma}^2_{03}(\hat{\theta}_i)$ represent the estimator of $\sigma^2(\hat{\theta}_i)$ based on U-statistic estimators of θ_i, p_{2i}, and p_{3i} for data from the i*th* investigator [see (A2) and (A3) of the Appendix]. The test of H_{03}: $\Sigma c_i(\theta_i-\frac{1}{2})=0$ is then based on comparing

$$V_3 = \frac{\sum_{i=1}^{a} c_i(\hat{\theta}_i-\frac{1}{2})}{\left[\sum_{i=1}^{a} c_i^2 \hat{\sigma}^2_{03}(\hat{\theta}_i)\right]^{\frac{1}{2}}}$$

to the standard normal distribution.

Recall that the usual van Elteren statistic V_E has the numerator of V_3 with $c_i=m_in_i/[m_i+n_i+1]$ and denominator variance based on the assumption that a null $\theta_i=\frac{1}{2}$ situation exists for each investigator. Typically V_E compared to a standard normal will be a conservative procedure under (3.1) since the null Wilcoxon variance $\sigma_0^2(\hat{\theta}_i)$ is usually larger than the non-null $\theta_i\neq\frac{1}{2}$ variance. (This can be seen by computing Var $\hat{\theta}_i$ for various parametric families, e.g., Lehmann, 1975, p. 400-401, or by computing the Birnbaum and Klose (1957) bounds on Var $\hat{\theta}_i$ under stochastically increasing alternatives. However, Hochberg, 1981, p. 1730, gives a counter-example in the ordered categorical data context). If m_i and n_i are small, the denominator of V_3 may be unstable and a more valid procedure may be to replace $\hat{\sigma}_{03}^2(\hat{\theta}_i)$ by $\sigma_0^2(\hat{\theta}_i)$ as in V_E.

To test the null hypothesis of no interaction H_{04}: $\theta_1=\ldots=\theta_a$ in the fixed effects situation, the procedure is the same as that for the mixed model (see Section 2.4). That is, we compare $V_2=\Sigma_{i=1}^a w_i(\hat{\theta}_i-\hat{\theta}_w)^2$ to the χ_{a-1}^2 distribution, where $w_i^{-1}=\hat{\sigma}_{02}^2(\hat{\theta}_i)$ and $\hat{\theta}_w$ are defined in Section 2.4. For ordered categorical data, there is a dependence on i of θ_i and other parameters in (A4), due to the fact that the block effect α_i does not cancel in the comparisons $I\big(g(Y_{ik})>g(X_{ij})\big)=I\big(g(\mu+\alpha_i+\beta_2+\epsilon_{yik})>g(\mu+\alpha_i+\beta_1+\epsilon_{xij})\big)$. Nevertheless, we feel that the dependence is small and still recommend V_2 for the test of interaction in the fixed effects situation.

3.2 Comments on the van Elteren and Pseudohomogeneity Statistics.

We have relied heavily on analogies with standard ANOVA methodology to motivate and develop our test procedures. Continuing in this spirit, we examine alternative methods currently used in analysis of data from multisite clinical trials in light of their relationship to ANOVA tests and hypotheses. Following Fleiss (1981, p. 162) for the *ith* investigator, let y_i represent an estimator of some measure of the difference between drugs. We assume y_i has

"large sample" expectation $\mu_i=0$, i=1, $\cdots$, a under the null hypothesis of no difference between drugs. For dichotomous responses y_i might be the log odds ratio but for our purposes y_i will be $\hat{\Delta}_i$ or $\hat{\theta}_i-\frac{1}{2}$. Also let $\tilde{y}=\Sigma v_i y_i/\Sigma v_i$, where v_i^{-1} is a consistent estimator of the variance of y_i possibly under a general alternative such as (3.1).

Certain commonly used test procedures (see, e.g., Fleiss 1981, p. 162) are based on the partitioning of an overall measure of drug effect given by

$$\chi^2_{\text{total}}=\chi^2_{\text{assoc}}+\chi^2_{\text{homog}}$$

or

$$\Sigma v_i y_i^2=\tilde{y}^2\Sigma v_i+\Sigma v_i(y_i-\tilde{y})^2. \tag{3.3}$$

If v_i^{-1} is a valid estimator of variance under a general alternative, this partitioning and the associated chi-squared tests are appropriate for the nested hypotheses

$$\{E(y_i)=0,\ i=1,\ \cdots,\ a\}\subset\{E(y_i)=\mu,\ i=1,\ \cdots,\ a\}\subset\{E(y_i)=\mu_i\}.$$

Defining main and interaction effects in terms of $E(y_i)$, the hypotheses can be written {no main effect, no interaction}$\subset${no interaction, possible main effect}$\subset${possible main effect and interaction}. If v_i^{-1} is an appropriate variance estimator under $\{E(y_i)=\mu,\ i=1,\ \ldots,\ a\}$, then χ^2_{homog} provides a test for interaction in the presence of a main effect. Often, however, v_i^{-1} is appropriate only under $\{E(y_i)=0,\ i=1,\ \ldots,\ a\}$, as in Q_{PH}, the pseudohomogeneity statistic in Koch and Edwards (1988, p. 418). The statistic χ^2_{assoc} is appropriate for testing $\{E(y_i)=0,\ i=1,\ \ldots,\ a\}$ against $\{E(y_i)=\mu,\ i=1,\ \ldots,\ a\}$. If interaction is present, χ^2_{assoc} focuses on the null $\Sigma v_i\mu_i=0$ rather than on $\Sigma\mu_i=0$ and so represents a test for a main effect only if interaction is absent (or the v_i are equal). Interpretation of the three tests is therefore not straightforward and description of these tests is often followed by lengthy discussion of their interpretation or implementation (e.g., Koch and Edwards, 1988, pp. 417-8, de Kroon and van der Laan, 1981, p. 195, and Fleiss, 1981, p. 164).

The difference between our proposed procedures and current methods based on the partitioning in (3.3) can be illustrated by appealing once more to ANOVA methodology. Let $y_i=\hat{\Delta}_i=\overline{Y}_i-\overline{X}_i$ and $v_i^{-1}=[\frac{1}{m_i}+\frac{1}{n_i}]\,\hat{\sigma}_\epsilon^2$ where $\hat{\sigma}_\epsilon^2$ is the Error Mean Square after fitting (3.1). Then (after factoring out $\hat{\sigma}_\epsilon^2$), (3.3) corresponds to the sequential or SAS Type I sums of squares due to fitting drug then investigator by drug after first fitting investigator. For balanced data (or $v_1=\cdots=v_a$) the hypotheses tested by $\chi^2_{assoc}=(\Sigma v_i\hat{\Delta}_i)^2/\Sigma v_i$ and $\chi^2_{homog}=\Sigma v_i(\hat{\Delta}_i-\tilde{\Delta})^2$ are, respectively, H_{01}: $\overline{\Delta}=0$ (no main effect) and H_{02}: $\Delta_1=\cdots=\Delta_a$ (no interaction). For unbalanced data, however, the null implied by $(\Sigma v_i\hat{\Delta}_i)^2/\Sigma v_i$ is not H_{01} and the ANOVA practitioner would without hesitation abandon the Type I partitioning of (3.3) and use instead the Type III sums of squares, or $F=(\frac{1}{a}\Sigma\hat{\Delta}_i)^2/(\frac{1}{a^2}\Sigma v_i^{-1})$, to test for a drug main effect.

The popular extended Mantel-Haenszel statistic Q_{EMH}, (Koch and Edwards, 1988, p. 420) is directly analogous to χ^2_{assoc} and hence to the SAS Type I or sequential partitioning in ANOVA. One version of Q_{EMH} is $(V_E)^2$, the square of the van Elteren statistic with "optimal" weights (χ^2_{assoc} with $y_i=\hat{\theta}_i-\frac{1}{2}$, $v_i=\frac{12n_im_i}{n_i+m_i+1}$). Given unbalanced data the ANOVA literature (e.g., Searle, 1987) strongly recommends testing meaningful hypotheses and we suggest the same philosophy should be applied to tests on the θ_i (or other measures of association such as the log odds ratio). This is accomplished using the tests we have introduced, rather than tests based on the partitioning (3.3).

4. A Multisite Clinical Trial.

The (slightly modified) data reported in Table 2 are from a clinical trial to compare a new drug with a placebo. A total of 9 investigators participated, each at a different site, and the number of subjects per investigator ranged from 6 to 22. Responses recorded were ordered

categorical in nature, with categories scored from 1 to 5, where 1 represents poorest performance and 5 the most improvement.

As explained in the Introduction (see also Section 5) it seems more appropriate to us to view investigator effects as random rather than as fixed, but for illustrative purposes we have carried out the analyses for both situations. Estimates $\hat{\theta}_i$ were therefore computed for each investigator and are presented in Table 3. The linear models (2.1) and (3.1) are not likely to be exactly correct for the highly discrete column scores (see, e.g., Harville and Mee, 1984), but for comparison with analyses of the $\hat{\theta}_i$, estimates $\hat{\Delta}_i$ were computed treating the column scores as measurements X and Y. Also for comparison and to provide useful insight, ANOVA was carried out on the column scores with the following results.

Source	df	Type III SS	F(with Inv*Drug as divisor)	F(with Error as divisor)
Investigator	8	8.239	0.80	1.75
Drug	1	7.634	5.94 (p=.041)	13.00 (p=.0005)
Investigator*Drug	8	10.290		2.19 (p=.033)
Error	113	66.352		

Due to unbalance and discreteness of the scores, these F tests are not exact but the overall impression is that there is evidence of an investigator by drug interaction and of a drug main effect. Treating investigator effects as random, the drug main effect should be tested against the investigator by drug interaction yielding the value F=5.94. Comparing F=5.94 with F=13.00 from the "fixed effects" test using the Error Mean Square as divisor suggests that the latter results in an overstatement of the drug main effect. These qualitative conclusions based on ANOVA are supported by the analyses of the $\hat{\theta}_i$ and of the $\hat{\Delta}_i$ discussed below. Note that all p values reported are for two sided alternatives so that comparisons with ANOVA F tests can easily be made.

To test for a drug main effect, assuming random investigator effects, our procedure is

essentially to compute a t statistic on the $\hat{\theta}_i - \frac{1}{2}$. In Table 3 this is seen to result in $V_1 = 2.26$ and a (two sided) p value of .054 based on comparison with the t_8 distribution. Also shown in Table 3 is the analogous test for a main effect based on assuming a linear model for the column scores. Computing a t statistic on the $\hat{\Delta}_i$ in Table 3 yields $V_\Delta = 2.53$ and a p value of .035. The smaller p value for V_Δ occurs because the few extreme responses (corresponding to column scores of 5) are given greater weight in computing mean differences $\hat{\Delta}_i$ than in the rank-based statistic V_1. P values for both V_1 and V_Δ are, however, reasonably close to that for the "mixed model" ANOVA approximate F test ($\sqrt{F_{1,8}} = \sqrt{5.94} = 2.44$, p=.041). The test for interaction yields $V_2 = 15.60$, p=.048. The data are ordered categorical so the appropriate variance estimator in the Appendix was used in computing V_2. That is, the quantities in (A4) were estimated separately for each investigator as in (A5), and these estimates were averaged across investigators using the weights d_i in Table 3. Then $\hat{\sigma}^2_{02}(\hat{\theta}_i)$ was computed from the averages using equation (A2). The resulting values are presented in Table 3. Comparing results for V_2 to the "mixed model" ANOVA F test for interaction based on the scores we see that both yield p values just less than .05. We conclude that there is reason to expect an overall effect for the new drug if used by a larger population of clinicians, and that there is evidence of some difference in effect for the new drug across clinicians.

The "fixed effects" analysis of the $\hat{\theta}_i$ is now described. The test for a main effect in the θ_i (i.e. for H_{03}: $\bar{\theta} = \frac{1}{2}$) is based on comparing V_3 with $c_i = 1/a = 1/9$ to the standard normal distribution. This results in $V_3 = 4.48$, p=.0000. For comparison, the "test for average partial association" based on $V_E = [\Sigma c_i(\hat{\theta}_i - \frac{1}{2})]/[\Sigma c_i^2 \sigma_0^2(\hat{\theta}_i)]^{\frac{1}{2}}$ and the standard normal results in $V_E = 3.56$, p=.0004 (see Table 3 for the c_i values). The difference between V_E and V_3 for these data is largely due to the difference between the variance estimators $\sigma^2_0(\hat{\theta}_i)$ and $\hat{\sigma}^2_{03}(\hat{\theta}_i)$. Although both of these estimators allow for discreteness of the data, the values listed in Table

3 show the "null" variance σ_0^2 is considerably larger than the estimated $\hat{\sigma}_{03}^2$, which does not assume $\theta_i=\frac{1}{2}$. This is probably due both to a tendency for σ_0^2 to be upwardly biased when $\theta_i \neq \frac{1}{2}$ (the bias increasing as θ_i approaches 0 or 1) and also to inaccuracy of the estimator $\hat{\sigma}_{03}^2$ when m_i, n_i are small. In spite of the differences between V_E and V_3, however, both yield p values that are close to that for the "fixed effects" ANOVA F test on the column scores ($F_{1,113}=13.00$, $p=.0005$). Our test for interaction in the "fixed effects" situation (as in the mixed model situation) is based on $V_2=15.60$, $p=.048$ (from the χ_8^2 distribution). The pseudohomogeneity statistic Q_{PH} (Koch and Edwards, 1988, p. 418) is computed by subtracting $(V_E)^2$ from the "total partial association" statistic. As shown in Table 3, $\chi_{total}^2=27.545$ yielding $Q_{PH}=27.545-(3.56)^2=14.880$, $p=.062$ (from the χ_8^2 distribution).

To summarize, we see that there is close agreement between our procedures and the analogous ANOVA tests, although they are aimed at different measures of effect. Also, there is generally good agreement between the current methods based on V_E and Q_{PH} and our "fixed effects" analysis of the $\hat{\theta}_i$. Note, however, the considerable difference between the tests for a drug main effect depending on whether the mixed model or fixed effects situation is assumed. The borderline p values for the mixed model tests (V_1, V_Δ, and ANOVA) are greater by about 2 orders of magnitude than those for the fixed effects tests (V_3, V_E). This clearly illustrates the potential for arriving at different conclusions depending on the test procedure used. In the next Section we discuss some of the reasons for preferring the mixed model approach.

5. Concluding Remarks

In discussing analysis of data from clinical trials, many authors (e.g., Koch and Edwards, 1988, p. 403, and references therein) suggest that the selection of subjects is not at random from some larger patient population. These authors note that an advantage of the extended

Mantel-Haenszel procedures including V_E is that they require only random allocation of subjects to treatments (and not random selection of subjects) but that a consequent disadvantage is that inferences must be limited to the study participants. As investigators (or sites) are not selected randomly either, it may be argued that our mixed model analysis assuming random investigator effects is inappropriate and that the resulting inferences are not valid. We counter such an argument with two observations. First, in spite of disclaimers that inferences should be limited to study participants, "randomization" p values for V_E are often interpreted in the broader sense of providing information about a potential drug main effect beyond the study population. Second, the bias introduced by the nonrandom selection of participating investigators will often be in the direction of underestimating the drug by investigator variance component, and carefully controlled treatment protocols will contribute to further underestimation of this interaction. The mixed model analysis will therefore be liberal because underestimation of the drug by investigator interaction will tend to cause overestimation of the drug main effect, <u>but</u> the mixed model analysis should be less liberal (or biased) than procedures such as V_E which ignore this interaction variance component entirely. In this type of study we believe that the p value for the mixed model test will be more realistic in the sense of predicting a drug main effect in larger and more heterogeneous populations of physicians and patients than the "sample specific" or "randomization" p value of the V_E test.

APPENDIX

Here we give basic formulas for the variance of $\hat{\theta}_i$ and U-statistic estimators of that variance for both continuous and ordered categorical data. These are needed for the tests of H_{02}, H_{03}, and H_{04} in Sections 2.4 and 3.1. For simplicity we suppress all notation having to do with investigators by focusing on data for a single investigator. Thus we assume throughout that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are iid with distribution functions $F(x)$ and $G(x)$, respectively.

A two-sample U-statistic of degree (r,s) with kernel $h(x_1, \ldots, x_r; y_1, \ldots, y_s)$ has the form (see Randles and Wolfe, 1979, p. 90-91)

$$U=\frac{1}{\binom{m}{r}\binom{n}{s}} \sum\sum h(x_{i_1}, \ldots, x_{i_r}; y_{j_1}, \ldots, y_{j_s}), \tag{A1}$$

where the sum is over all subsets of size r and s from the respective samples. When (r,s)=(1,1) the variance of U is

$$\text{Var } U=\frac{1}{mn}\left[(m-1)\left(\gamma_{0,1}-(EU)^2\right)+(n-1)\left(\gamma_{1,0}-(EU)^2\right)+\gamma_{1,1}-(EU)^2\right], \tag{A2}$$

where $\gamma_{0,1}=Eh(X_1;Y_1)h(X_2;Y_1)$, $\gamma_{1,0}=Eh(X_1;Y_1)h(X_1;Y_2)$, and $\gamma_{1,1}=E[h(X_1;Y_1)]^2$. Since under H_{02}, H_{03}, or H_{04}, (X_1, Y_1) and (X_2, Y_2) are independent, (A2) does not contain a term corresponding to the final term in (2.2).

For continuous data, $\hat{\theta}$ has kernel $h(x;y)=I(y>x)$ and $\gamma_{1,0}=p_2=P(Y_1>X_1, Y_2>X_1)$, $\gamma_{0,1}=p_3=P(Y_1>X_1, Y_1>X_2)$, and $\gamma_{1,1}=\theta^2$. The U-statistic estimators of p_2 and p_3 are from (A1)

$$\hat{p}_2=\frac{2}{mn(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} I(Y_j>X_i, Y_k>X_i)$$

$$\hat{p}_3 = \frac{2}{m(m-1)n} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=1}^{n} I(Y_k > X_i,\ Y_k > X_j).$$

We can give simple expressions for $\hat{p}_2$ and $\hat{p}_3$ in terms of ranks by first ordering the samples $X_{(1)} \le \ldots \le X_{(m)}$ and $Y_{(1)} \le \ldots \le Y_{(n)}$. If $Q_{(i)}$ and $R_{(j)}$ are the ranks of $X_{(i)}$ and $Y_{(j)}$ in the combined data, then

$$\hat{p}_2 = \frac{1}{mn(n-1)} \sum_{i=1}^{m} \left[n+i-Q_{(i)}\right]^2 - \frac{\hat{\theta}}{n-1}$$

$$\hat{p}_3 = \frac{1}{m(m-1)n} \sum_{j=1}^{n} \left[m+j-R_{(j)}\right]^2 - \frac{\hat{\theta}}{m-1}. \qquad \text{(A3)}$$

Our purpose here was to get U-statistic estimators of the components of Var $\hat{\theta}$ so that they could be averaged over different investigators and the averages substituted in (A2). For estimating Var $\hat{\theta}$ from a single investigator, we could of course use (A2) with EU replaced by $\hat{\theta}$, $\gamma_{1,0}$ replaced by $\hat{p}_2$, and $\gamma_{0,1}$ replaced by $\hat{p}_3$. If there are just a few ties with generally continuous data, we suggest using (A3) with mid-ranks.

For ordered categorical data, the U-statistic theory is still appropriate but our definition of θ changes so that $\hat{\theta}$ has kernel $h(x;y) = I(y>x) + \frac{1}{2}I(y=x)$. This kernel leads to Var $\hat{\theta}$ given by (A2) with

$$\begin{aligned} \gamma_{1,0} &= p_2 + P(Y_1 > Y_2 = X_1) + \tfrac{1}{4}P(Y_1 = Y_2 = X_1), \\ \gamma_{0,1} &= p_3 + P(Y_1 = X_2 > X_1) + \tfrac{1}{4}P(Y_1 = X_1 = X_2), \\ \gamma_{1,1} &= \theta - \tfrac{1}{4}P(Y_1 = X_1). \end{aligned} \qquad \text{(A4)}$$

Suppose that from the X and Y samples we construct the contingency table

	1	2	3	...	t	
X's	f_1	f_2	f_3	...	f_t	$m=\sum f_i$
Y's	g_1	g_2	g_3	...	g_t	$n=\sum g_i$

with frequencies f_i and g_i for ordered categories $i=1, \ldots, t$. Then using (A1) reexpressed in the frequency notation gives

$$\hat{\gamma}_{1,0}=\hat{p}_2+\frac{1}{mn(n-1)}\left[\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} f_i g_i g_j+\frac{1}{4}\sum_{i=1}^{t} f_i g_i(g_i-1)\right]$$

$$\hat{\gamma}_{0,1}=\hat{p}_3+\frac{1}{m(m-1)n}\left[\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} f_i f_j g_j+\frac{1}{4}\sum_{i=1}^{t} f_i(f_i-1)g_i\right]$$

$$\hat{\gamma}_{1,1}=\hat{\theta}-\frac{1}{4mn}\sum_{i=1}^{t} f_i g_i \qquad \text{(A5)}$$

$$\hat{\theta}=\frac{1}{mn}\left[\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} f_i g_j+\frac{1}{2}\sum_{i=1}^{t} f_i g_i\right]$$

$$\hat{p}_2=\frac{1}{mn(n-1)}\left[2\sum_{i=1}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} f_i g_j g_k+\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} f_i g_j(g_j-1)\right]$$

$$\hat{p}_3=\frac{1}{m(m-1)n}\left[2\sum_{i=1}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} f_i f_j g_k+\sum_{i=1}^{t-1}\sum_{j=i+1}^{t} f_i(f_i-1)g_j\right].$$

For comparison note that the estimate of Var $\hat{\theta}$ obtained by using (A2) with these estimators is $(2mn)^{-2}\hat{V}_{F,G}(W)$, where $\hat{V}_{F,G}(W)$ is given by Hochberg (1981, p. 1726, and correcting the errors in his $\hat{\Pi}_{xyy}$, $\hat{\Pi}_{yxy}$, and $\hat{\Pi}_{yyx}$ by replacing $m-1$ with $n-1$).

REFERENCES

Birnbaum, Z.W. and Klose, O.M. (1957). Bounds for the variance of the Mann-Whitney statistic. *Annals of Mathematical Statistics* 38, 933-945.

de Kroon, J. and van der Laan, P. (1981). Distribution-free test procedures in two-way layouts; a concept of rank-interaction. *Statistica Neerlandica* 35, 189-213.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* 2nd Edition. New York: Wiley.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41, 361-372.

Gupta, S.S. (1963). Probability integrals of multivariate normal and multivariate t. *Annals of Mathematical Statistics* 34, 792-828.

Harville, D.A. and Mee, R.W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics* 40, 393-408.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks.* New York: Wiley.

Hochberg, Y. (1981). On the variance estimate of a Wilcoxon-Mann Whitney statistic for group ordered data. *Communications in Statistics-Theory and Methods* A 10 (17), 1719-1732.

Koch, G.G. and Edwards, S. (1988). Clinical efficacy trials with categorical data. In *Biopharmaceutical Statistics for Drug Development*, K.E. Peace (ed.), 403-457. New York: Marcel Dekker.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

Searle, S.R. (1987). *Linear Models for Unbalanced Data.* New York: Wiley.

van Elteren, P.H. (1960). On the combination of independent two sample tests of Wilcoxon. *Bulletin of the International Statistical Institute* 37, 351-361.

Webster, J.T. (1968). An approximate F-statistic. *Technometrics* 10, 597-604.

Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association* 29, 51-66.

Table 1

Minimum Sample Sizes for Var $\hat{\theta}_i$ and Var $\hat{\Delta}_i$ to exceed their asymptotic values $(p_{4i}-\frac{1}{4})$ and $2\sigma_\gamma^2$ by no more than 25%.

$$\text{Var}\ \hat{\theta}_i/(p_{4i}-\tfrac{1}{4}) \leq 1.25$$

	$\sigma_\gamma^2/\sigma_\epsilon^2=\frac{1}{4}$		$\sigma_\gamma^2/\sigma_e^2=1$		$\sigma_\gamma^2/\sigma_\epsilon^2=4$	
	n_i	m_i	n_i	m_i	n_i	m_i
$n_i=m_i$	19	19	6	6	3	3
$3n_i=m_i$	12	36	4	12	2	6

$$\text{Var}\ \hat{\Delta}_i/2\sigma_\gamma^2 \leq 1.25$$

$n_i=m_i$	16	16	4	4	1	1
$3n_i=m_i$	11	33	3	9	1	3

Note: Results for Var $\hat{\theta}_i$ are for normal data.

Table 2

Frequency distribution of subject responses from a multisite clinical trial to compare a new drug with a placebo.

Investigator	Drug	Score 1	2	3	4	5	Total number of subjects
1	Placebo	1	3	4	2	0	10
	New	0	1	5	4	0	10
2	Placebo	0	2	2	1	0	5
	New	0	1	2	1	0	4
3	Placebo	0	3	2	3	0	8
	New	0	0	0	5	3	8
4	Placebo	0	1	5	3	0	9
	New	0	2	3	3	0	8
5	Placebo	0	3	3	5	0	11
	New	0	3	4	3	1	11
6	Placebo	0	1	6	0	0	7
	New	0	1	4	1	2	8
7	Placebo	0	1	2	0	0	3
	New	0	2	1	0	0	3
8	Placebo	0	4	4	0	0	8
	New	0	0	2	6	0	8
9	Placebo	0	2	4	0	0	6
	New	0	0	1	3	0	4

Table 3

Estimates of the drug performance measures θ_i and Δ_i for each investigator, variance estimates, and test statistics for assessing drug main and interaction effects.

	$\hat{\Delta}_i$	$\hat{\theta}_i-\frac{1}{2}$	$c_i=\frac{n_i m_i}{n_i+m_i+1}$	$d_i=\frac{c_i}{\Sigma c_i}$	$\hat{\sigma}^2_{02}(\hat{\theta}_i)$	$\hat{\sigma}^2_{03}(\hat{\theta}_i)$	$\sigma^2_0(\hat{\theta}_i)$	$\chi^2_i=\frac{(\hat{\theta}_i-\frac{1}{2})^2}{\sigma^2_0(\hat{\theta}_i)}$
1	0.60	.185	4.76	.156	.014745	.010572	.015329	2.232
2	0.20	.075	2.00	.066	.033469	.023375	.036458	0.154
3	1.38	.383	3.76	.123	.018429	.003190	.019108	7.669
4	−0.10	−.028	4.00	.131	.017494	.015164	.017693	0.044
5	0.00	−.012	5.26	.172	.013406	.012387	.014275	0.011
6	0.64	.170	3.50	.115	.019630	.011061	.016709	1.722
	−0.33	−.167	1.29	.047	.049076	.024691	.050000	0.556
8	1.25	.438	3.76	.123	.018429	.001526	.019531	9.800
9	1.08	.417	2.18	.071	.031231	.004340	.032408	5.357

$\bar{\hat{\Delta}}=.525$ $\bar{\hat{\theta}}-\frac{1}{2}=.162$

$V_\Delta=2.53$ $V_1=2.26$ $V_2=15.60$ $V_3=4.48$ $V_E=3.56$ $\chi^2_{total}=27.55$

Note: The variance estimate $\hat{\sigma}^2_{03}(\hat{\theta}_i)$, is obtained from (A2) with $m=m_i$, $n=n_i$ and using estimates $(\hat{\theta}, \hat{\gamma}_{1,0}, \hat{\gamma}_{0,1}, \hat{\gamma}_{1,1})$ in (A5) from data for the i^{th} investigator only. The variance estimate $\hat{\sigma}^2_{02}(\hat{\theta}_i)$ is also obtained from (A2) with $m=m_i$, $n=n_i$, but using estimates of $(\theta, \gamma_{1,0}, \gamma_{0,1}, \gamma_{1,1})$ obtained by averaging individual $(\hat{\theta}, \hat{\gamma}_{1,0}, \hat{\gamma}_{0,1}, \hat{\gamma}_{1,1})$ across investigators using the weights d_i in column 4. Finally, $\sigma^2_0(\hat{\theta}_i)$ is the "null" variance adjusted for ties (Lehmann, 1975, p.20).