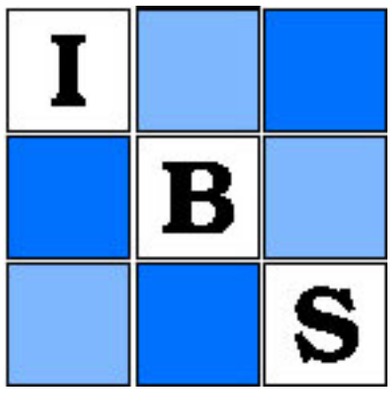

The Analysis of Multiple Endpoints in Clinical Trials
Author(s): Stuart J. Pocock, Nancy L. Geller and Anastasios A. Tsiatis
Source: *Biometrics*, Vol. 43, No. 3 (Sep., 1987), pp. 487-498
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/2531989
Accessed: 05/08/2013 17:51

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

*International Biometric Society* is collaborating with JSTOR to digitize, preserve and extend access to *Biometrics.*

http://www.jstor.org

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

BIOMETRICS 43, 487–498
September 1987

# The Analysis of Multiple Endpoints in Clinical Trials

**Stuart J. Pocock,[1] Nancy L. Geller,[2] and Anastasios A. Tsiatis[3]**

[1] Department of Clinical Epidemiology and General Practice,
Royal Free Hospital School of Medicine, University of London,
London NW3 2PN, United Kingdom.
[2] Biostatistics, Memorial Sloan-Kettering Cancer Center,
New York, New York 10021, U.S.A.
[3] Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts 02115, U.S.A.

SUMMARY

Treatment comparisons in randomized clinical trials usually involve several endpoints such that conventional significance testing can seriously inflate the overall Type I error rate. One option is to select a single primary endpoint for formal statistical inference, but this is not always feasible. Another approach is to apply Bonferroni correction (i.e., multiply each *P*-value by the total number of endpoints). Its conservatism for correlated endpoints is examined for multivariate normal data. A third approach is to derive an appropriate global test statistic and this paper explores one such test applicable to any set of asymptotically normal test statistics. Quantitative, binary, and survival endpoints are all considered within this general framework. Two examples are presented and the relative merits of the proposed strategies are discussed.

## 1. Introduction

The excessive use of multiple significance tests in clinical trials can greatly increase the chance of false positive findings. This paper addresses the problem of how to apply significance testing in randomized trials with several outcome measures for treatment comparison. The problem is complicated by the fact that endpoints are usually correlated and trials often have a mixture of data types, e.g., quantitative, binary, and survival data.

Perhaps the most common approach in the medical literature is to analyze each endpoint separately, presenting multiple *P*-values and an overall subjective conclusion. At best, this provides an open display of data enabling readers to draw their own (possibly different) conclusions. At worst, authors may emphasize significant findings, perhaps not even reporting some nonsignificant endpoints. Even if all endpoints are reported openly, authors and readers may still not appreciate the increased risk of an overall Type I error rate. It is commonplace to interpret a trial as positive if any endpoint has a treatment difference significant at the 5% level, and there is a need to deter authors from such indiscriminate use of *P*-values.

One possible solution is to specify in the study protocol a *single primary endpoint* whose *P*-value for treatment difference represents a formal test of hypothesis. All other endpoints are then subsidiary, requiring exploratory rather than formal interpretation. This simplified situation can be hard to maintain. For example, in heart disease prevention trials both nonfatal myocardial infarctions and total deaths are of interest, so emphasis of one over the other is judgmental. Also, there is often pressure (e.g., from journal editors and referees)

*Key words:* Clinical trial analysis; Multiple endpoints; Multiplicity.



This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

to provide *P*-values for subsidiary endpoints. Thus, selection of a single primary endpoint is only a partial solution to the multiple-endpoints problem, and it may be wise to prespecify more stringent Type I error rates for secondary endpoints.

This paper considers how to control *P*-values when all endpoints are analyzed on equal terms. For several normally distributed endpoints, one standard approach is Hotelling's $T^2$ (Press, 1972). However, as pointed out by O'Brien (1984) and Meier (1975), Hotelling's $T^2$ is intended to detect *any* departure from the null hypothesis and hence lacks power to detect any specific types of departure that are considered a priori to be biologically plausible. Thus, Hotelling's $T^2$ is quite unsuitable for analysis of clinical trials and is not considered further.

For $k$ endpoints without prespecified priorities, how can significance testing be used while (i) preserving a small overall Type I error rate and (ii) allowing for correlated endpoints? Section 2 considers the conservatism of Bonferroni correction for *P*-values while Section 3 explores a global test statistic for any set of asymptotically normal test statistics, with particular reference to a method proposed by O'Brien (1984). Section 4 presents two examples and Section 5 discusses the relative merits of the alternative approaches.

## 2. Bonferroni Correction

The Bonferroni inequality can be used to obtain an adjustment to the smallest *P*-value for significance tests on $k$ endpoints (Miller, 1981; Armitage and Parmar, 1986). If the $k$ endpoints are independent then $\Pr(\text{smallest } P\text{-value} \leq \alpha) = 1 - (1 - \alpha)^k \approx \alpha k$ if $\alpha$ is small. Hence, Bonferroni correction has each *P*-value multiplied by $k$, the number of endpoints. That is, for an overall Type I error rate $\alpha$, one accepts as statistically significant only those *P*-values less than $\alpha/k$. In practice, endpoints are correlated, so that Bonferroni correction becomes conservative and Worsley (1982) has proposed one possible improvement. However, does such overcorrection seriously affect the power of a trial to detect genuine treatment differences?

It is difficult to obtain general results for nonnormal data or arbitrary correlations between endpoints. Therefore, we consider $k$ normally distributed endpoints, each with known variance, for which all possible pairs have the same (known) correlation $\rho$ within each of two treatment groups. Then, for any prespecified $\alpha$, $k$, and $\rho$ one can derive numerically that "nominal" value $\alpha'$ which the smallest of $k$ one-sided *P*-values obtained from the normal test statistics will reach with probability $\alpha$ under the null hypothesis. Table 1 shows values of $\alpha'$ and its corresponding standardized normal deviate $z'$ for $k = 2, \ldots, 10$ endpoints; for $\rho = 0, .1, .3, .5, .7, .9$; and for $\alpha = .05, .025$. Values of $z'$ were obtained using a quintic interpolation of tabular values for the probability distribution of the maximum of $k$ equicorrelated standardized normal deviates published by Gupta (1963). Table 1 also shows the exact values of $\alpha'$ for $\rho = 0$, which are slightly smaller than $\alpha/k$.

For each number of endpoints $k$, $\alpha'$ increases as the correlation between endpoints increases, i.e., the conservatism of Bonferroni increases as $\rho$ increases. However, the degree of conservatism is small for $\rho \leq .5$. For instance, with $k = 5$ endpoints, $\alpha = .05$, and $\rho = .5$, $\alpha' = .0128$, compared with $\alpha/k = .01$. Thus, Bonferroni correction works reasonably well for moderately correlated variables. Also, there is no noticeable deterioration in Bonferroni correction as the number of correlated endpoints increases. For two-sided testing, approximate results are obtained by doubling every one-sided probability (both for $\alpha$ and $\alpha'$) in Table 1. Thus, for two-sided $\alpha = .05$ one can use one-sided $\alpha = .025$. The results are very slightly conservative, since the probability of any two variables exceeding the critical $\alpha'$ levels in opposite directions is ignored, but this is negligibly small.

In reality, multiple endpoints are not usually equicorrelated and normally distributed. However, it seems plausible that similar findings would occur for any continuous

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

**Table 1**

*Bonferroni correction and the true "nominal" significance levels $\alpha'$, with corresponding standardized normal deviates $z'$ to preserve an overall one-sided Type I error $\alpha$, for $k = 2(1)10$ normally distributed endpoints and equal pairwise correlations $\rho = 0, .1(.2).9$ within each of two treatment groups*

| | | | $\rho$ | | | | |
|---|---|---|---|---|---|---|---|
| $k$ | Bonferroni | 0 | .1 | .3 | .5 | .7 | .9 |
| **$\alpha = .05$** | | | | | | | |
| 2 | $z'$ 1.960 | 1.955 | 1.951 | 1.938 | 1.916 | 1.877 | 1.798 |
| | $\alpha'$ .0250 | .0253 | .0255 | .0263 | .0277 | .0303 | .0361 |
| 3 | $z'$ 2.128 | 2.121 | 2.116 | 2.097 | 2.064 | 2.001 | 1.877 |
| | $\alpha'$ .0167 | .0170 | .0172 | .0180 | .0195 | .0227 | .0303 |
| 4 | $z'$ 2.241 | 2.234 | 2.228 | 2.204 | 2.160 | 2.083 | 1.929 |
| | $\alpha'$ .0125 | .0127 | .0130 | .0138 | .0154 | .0186 | .0269 |
| 5 | $z'$ 2.326 | 2.319 | 2.312 | 2.285 | 2.233 | 2.144 | 1.967 |
| | $\alpha'$ .0100 | .0102 | .0104 | .0112 | .0128 | .0160 | .0246 |
| 6 | $z'$ 2.394 | 2.386 | 2.379 | 2.349 | 2.290 | 2.193 | 1.998 |
| | $\alpha'$ .0083 | .0085 | .0087 | .0094 | .0110 | .0142 | .0229 |
| 7 | $z'$ 2.450 | 2.442 | 2.434 | 2.402 | 2.340 | 2.233 | 2.022 |
| | $\alpha'$ .0071 | .0073 | .0075 | .0082 | .0096 | .0128 | .0216 |
| 8 | $z'$ 2.498 | 2.490 | 2.489 | 2.447 | 2.381 | 2.267 | 2.043 |
| | $\alpha'$ .0063 | .0064 | .0064 | .0072 | .0086 | .0117 | .0205 |
| 9 | $z'$ 2.539 | 2.531 | 2.522 | 2.487 | 2.417 | 2.296 | 2.062 |
| | $\alpha'$ .0056 | .0057 | .0058 | .0065 | .0079 | .0108 | .0196 |
| 10 | $z'$ 2.576 | 2.568 | 2.559 | 2.521 | 2.448 | 2.322 | 2.077 |
| | $\alpha'$ .0050 | .0051 | .0053 | .0059 | .0073 | .0101 | .0189 |
| **$\alpha = .025$** | | | | | | | |
| 2 | $z'$ 2.241 | 2.240 | 2.237 | 2.229 | 2.212 | 2.180 | 2.108 |
| | $\alpha'$ .0125 | .0126 | .0127 | .0129 | .0135 | .0146 | .0175 |
| 3 | $z'$ 2.394 | 2.391 | 2.388 | 2.375 | 2.350 | 2.298 | 2.185 |
| | $\alpha'$ .0083 | .0084 | .0085 | .0088 | .0094 | .0108 | .0144 |
| 4 | $z'$ 2.498 | 2.494 | 2.491 | 2.475 | 2.442 | 2.377 | 2.237 |
| | $\alpha'$ .0063 | .0063 | .0064 | .0067 | .0073 | .0087 | .0127 |
| 5 | $z'$ 2.576 | 2.572 | 2.568 | 2.550 | 2.511 | 2.436 | 2.274 |
| | $\alpha'$ .0050 | .0051 | .0051 | .0054 | .0060 | .0074 | .0115 |
| 6 | $z'$ 2.638 | 2.635 | 2.630 | 2.611 | 2.567 | 2.483 | 2.304 |
| | $\alpha'$ .0042 | .0042 | .0043 | .0045 | .0051 | .0065 | .0106 |
| 7 | $z'$ 2.690 | 2.687 | 2.682 | 2.661 | 2.613 | 2.521 | 2.328 |
| | $\alpha'$ .0036 | .0036 | .0037 | .0039 | .0045 | .0058 | .0100 |
| 8 | $z'$ 2.734 | 2.731 | 2.729 | 2.703 | 2.652 | 2.554 | 2.349 |
| | $\alpha'$ .0031 | .0032 | .0032 | .0034 | .0040 | .0053 | .0094 |
| 9 | $z'$ 2.773 | 2.769 | 2.764 | 2.740 | 2.686 | 2.583 | 2.367 |
| | $\alpha'$ .0028 | .0028 | .0029 | .0031 | .0036 | .0049 | .0090 |
| 10 | $z'$ 2.807 | 2.803 | 2.798 | 2.773 | 2.716 | 2.608 | 2.383 |
| | $\alpha'$ .0025 | .0025 | .0026 | .0028 | .0033 | .0047 | .0086 |

asymptotically normal test statistics, though for discrete data there will be an inevitable degree of conservatism. For unequally correlated variables, it is difficult to give any generalizable results. However, if most pairwise correlations are less than .5, serious conservatism should not occur. If any two variables are known to be highly correlated ($\rho = .9$, say) it would be sensible to preselect one of them or to predefine some combination of them.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

The main drawback to Bonferroni correction is that it confines attention to the smallest $P$-value of $k$ test statistics. Thus, five endpoints with $P$-values of .01, .7, .7, .7, .7 are considered more highly significant than five endpoints all at $P = .02$, whereas the latter appears to contain more convincing evidence of a treatment difference.

Thus, Bonferroni correction has its greatest power for alternative hypotheses in which only one of $k$ endpoints has a nonzero treatment difference and, furthermore, one does not know in advance which endpoint that will be—a situation unlikely to arise in practice. For alternative hypotheses in which several variables depart from zero treatment difference in the same direction, Bonferroni correction will seriously lack power. However, since it is simple to apply, only slightly conservative, and easily understood, Bonferroni correction is still useful where the situation does not warrant more complex procedures.

## 3. A Global Test Statistic

Consider a randomized clinical trial with two treatment groups and $k$ correlated endpoints. Later in this section we study the general problem of $k$ endpoints with asymptotically normal test statistics, including binary data and survival data, but we begin with the multivariate normal case.

Often one has several quantitative endpoints that are biologically related and positively correlated. Prime interest is in alternative hypotheses with all (or some) endpoints showing treatment differences in the same direction. O'Brien (1984) considers this problem and here extensions of his parametric (generalized least squares) test statistic are provided for the two-sample problem.

First assume the $k$ endpoints have a multivariate normal distribution with known variances and known correlation matrix $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})$ in each of two treatment groups. Intuitively, an appropriate test statistic should be a linear combination of the univariate standardized normal deviates $z_1, \ldots, z_k$. Using generalized least squares, O'Brien defines the coefficients in this linear combination for an optimal test against the alternative hypothesis that the $k$ standardized treatment differences are all of equal magnitude and in the same direction. Let $\mathbf{J}' = (1, 1, \ldots, 1)$. Then O'Brien's test statistic may be written as

$$\frac{\mathbf{J}'\boldsymbol{\Sigma}^{-1}\mathbf{z}}{(\mathbf{J}'\boldsymbol{\Sigma}^{-1}\mathbf{J})^{1/2}} \tag{1}$$

which has a standardized normal distribution under the null hypothesis. The weighting factors $\mathbf{J}'\boldsymbol{\Sigma}^{-1}$ are in fact the column sums of $\boldsymbol{\Sigma}^{-1}$ for each variable, so endpoints that are less highly correlated with the other variables have correspondingly greater weights. Note that $\boldsymbol{\Sigma}$, the correlation matrix for the raw data in each treatment group, is also the correlation matrix for the standardized normal deviates $z_1, \ldots, z_k$. Also, $\mathbf{J}'\boldsymbol{\Sigma}^{-1}\mathbf{J}$ is the sum of all cells of $\boldsymbol{\Sigma}^{-1}$.

To understand further this test statistic, consider the following two hypothetical examples.

*Example* 1 *All endpoints equally correlated*, i.e., $\boldsymbol{\Sigma}_{ij} = \rho$ for all $i \neq j$.

Then $\boldsymbol{\Sigma}^{-1}$ has diagonal elements $d = 1 - \rho^2(k-1)/[(k-1)\rho^2 - 1 - (k-2)\rho]$ and off-diagonal elements equal to $(1-d)/[\rho(k-1)]$. All endpoints are weighted equally and formula (1) becomes

$$\frac{\bar{z}}{\{[1 + (k-1)\rho]/k\}^{1/2}} \sim \mathrm{N}(0, 1)$$

where $\bar{z}$ is the mean of $z_1, z_2, \ldots, z_k$. Two-sided 5% significance is achieved if $\bar{z} > 1.96\{[1 + (k-1)\rho]/k\}^{1/2}$, which decreases with $k$ and increases with $\rho$. For $k = 5$ endpoints and $\rho = 0$, $\bar{z} > .88$ has $P < .05$. For $k = 5$ and $\rho = .5$, $\bar{z} > 1.52$ has $P < .05$.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

Thus, the combined evidence of several endpoints in the same direction obviously need not be as extreme as for a single endpoint.

*Example* 2 *Some endpoints independent, others equicorrelated*, i.e., for all $i \neq j$, $\Sigma_{ij} = 0$ if $i < m$ or $j < m$, and $\Sigma_{ij} = \rho$ otherwise.

Formula (1) then becomes

$$\frac{\sum_{i \leq m} z_i + (\sum_{i>m} z_i)/[(k - m - 1)\rho + 1]}{\{m + (k - m)/[(k - m - 1)\rho + 1]\}^{1/2}} \sim \mathrm{N}(0, 1).$$

The $m$ independent endpoints have larger weights than the $(k - m)$ equicorrelated endpoints. However, the sum of these $(k - m)$ correlated $z$-values will have greater weight than any single uncorrelated $z$-value.

### 3.1 *Adaptation to Any Asymptotically Normal Test Statistics*

Extension of the above global test statistic to any set of asymptotically normal test statistics whose correlation matrix can be estimated is now illustrated for several types of data.

(i) *Normal endpoints with unknown variance–covariance matrix which is the same for both treatment groups* O'Brien (1984) proposed replacing $\boldsymbol{\Sigma}$ by the usual pooled within-treatment estimate $\mathbf{S}$. $\mathbf{z}$ is then replaced by the $k$ two-sample $t$-statistics $\mathbf{t}$ and formula (1) becomes

$$\frac{\mathbf{J}'\mathbf{S}^{-1}\mathbf{t}}{(\mathbf{J}'\mathbf{S}^{-1}\mathbf{J})^{1/2}}. \tag{2}$$

This has an asymptotic standardized normal distribution, but does not follow a $t$ distribution. We have undertaken some simulation studies to supplement those already reported by O'Brien (1984). Briefly, for $k = 2$ endpoints a $t$ distribution on $N - 4$ degrees of freedom appears a good approximation even for $N$ quite small (e.g., $N = 10$). However, for $k = 5$ endpoints, convergence to a standardized normal distribution is slower. For instance, the 5% point of formula (2) falls below 2.0 only for $N$ around 200 patients and is around 2.2 for $N = 40$ patients.

(ii) *Normal endpoints with unequal variance–covariance matrices in the two treatment groups, containing $n_1$ and $n_2$ patients, respectively* Suppose $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ are the (known) correlation matrices for the raw data in each of the two treatment groups and let $\bar{\mathbf{x}}_1$, $\boldsymbol{\sigma}_1$, and $\bar{\mathbf{x}}_2$, $\boldsymbol{\sigma}_2$ be the sample mean vectors and (known) standard deviation vectors for the $k$ endpoints in the two treatments.

If we define $z_i = (\bar{x}_{i1} - \bar{x}_{i2})/[\sigma_{i1}^2/n_1 + \sigma_{i2}^2/n_2]^{1/2}$ for $i = 1, \ldots, k$ and $\boldsymbol{\Sigma}_{ij} = (n_1\boldsymbol{\Sigma}_{ij1} + n_2\boldsymbol{\Sigma}_{ij2})/(n_1 + n_2)$ for all $i \neq j$, then formula (1) is applicable as before. Usually, one needs to estimate $\sigma_1$, $\sigma_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ in the usual way, in which case formula (1) is asympotically normal.

(iii) *Crossover trials with normal endpoints* For the two-period crossover trial there are established methods of analyzing a single endpoint (Armitage and Hills, 1982). Suppose such a trial has $k$ quantitative endpoints and for simplicity assume there are no period or carryover effects. Then each patient's data may be summarized by the $k$ treatment differences $d_1, \ldots, d_k$. Suppose these are normally distributed with an estimated correlation matrix $\mathbf{S}$ obtained from the $N$ patients' data.

For the $k$ endpoints one can obtain values of paired $t$-test statistics, $t_1, \ldots, t_k$, each with $N - 1$ degrees of freedom. Then a global test statistic is

$$\frac{\mathbf{J}'\mathbf{S}^{-1}\mathbf{t}}{(\mathbf{J}'\mathbf{S}^{-1}\mathbf{J})^{1/2}}. \tag{3}$$

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

Again this has an asymptotic standardized normal distribution, but does not follow a $t$ distribution. Simulation studies show that for $k = 2$ endpoints a $t$ distribution on $N - 2$ degrees of freedom appears a good approximation even for $N$ as small as 10 patients. However, for $k = 5$ endpoints convergence to normality is slower. For instance, the 5% point of formula (3) reaches 2.0 for $N$ around 100 patients and is 2.17 for $N = 50$ patients.

(iv) *Nonnormal quantitative data* In practice, quantitative endpoints are not usually normally distributed. For a single endpoint the robustness of $t$ or $z$ statistics must be assessed, taking into account the degree of skewness and the sample size. With multiple endpoints the same principles should hold for a global test statistic.

(v) *Binary data* For trials with two or more binary endpoints, one can use the normal approximation to the binomial. Let $p_{i1}, p_{i2}$ be the proportions responding to each treatment for the $i$th binary response variable, let $n_1, n_2$ be the number of patients in each treatment group, and let $N = n_1 + n_2$. Define $\bar{p}_i = (p_{i1}n_1 + p_{i2}n_2)/N$. Then

$$z_i = \frac{p_{i1} - p_{i2}}{[(\bar{p}_i(1 - \bar{p}_i)N/n_1n_2]^{1/2}} \tag{4}$$

is asymptotically N(0, 1) under the null hypothesis. The correlation between $z_i$ and $z_j$, $\boldsymbol{\Sigma}_{ij}$, is estimated by maximum likelihood as

$$\frac{s_{ij} - \bar{p}_i\bar{p}_j}{[\bar{p}_i\bar{p}_j(1 - \bar{p}_i)(1 - \bar{p}_j)]^{1/2}} \tag{5}$$

for any $i \neq j$, where $s_{ij}$ is the proportion of all patients with responses for both variables $i$ and $j$. Then, substituting (5) for $\boldsymbol{\Sigma}_{ij}$ in $\boldsymbol{\Sigma}$ and using formula (1) provides an asymptotically normal global test statistic. Further research is needed to assess this approximation for small samples, though it seems reasonable to suppose that if the sample sizes are adequate for each univariate normal approximation $z_i$, then the global test should also be an adequate approximation. Also, the use of a continuity correction for each $z_i$ may be inappropriate, since the corrected global test would then be conservative, a common problem for attempts at continuity correction for combinations of discrete asymptotically normal test statistics. Estimates of $\boldsymbol{\Sigma}_{ij}$ can also be obtained for mixtures of binary and normal endpoints.

(vi) *Survival data using log-rank tests* Suppose endpoint $j$ is a censored variable—say, patient survival times. Then the normal approximation to the log-rank test is given by

$$z_j = \frac{(\sum \delta_{m1} - E_j)}{V_j^{1/2}}, \tag{6}$$

where $\delta_{m1} = 1$ if the $m$th subject on treatment 1 is dead, $\delta_{m1} = 0$ otherwise, and $E_j$ and $V_j$ are the usual expectation and variance estimates for the log-rank test.

Let us consider the bivariate problem of *combining log-rank and binary endpoints*, e.g., survival time and tumour response within a preset time for a two-armed cancer chemotherapy trial. Then $z_j$ is defined in equation (6), $z_i$ is defined in equation (4), and, based on arguments given in the Appendix, $\boldsymbol{\Sigma}_{ij}$ can be estimated by

$$\frac{R_d - \sum \bar{p}(t_j)}{N[NV_j\bar{p}_i(1 - \bar{p}_i)/n_1n_2]^{1/2}}, \tag{7}$$

where $R_d$ is the total number of patients who both responded and died, $\bar{p}(u)$ is the proportion of patients who responded among those at risk at time $u$, and the summation is over all death times $t_j$. To obtain an anticipated positive correlation, one can consider the proportion of patients failing to respond.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

For this survival problem, the asymptotic normality of formula (1) can be utilized. Further adaptations are envisaged for other combinations of asymptotically normal test statistics, e.g., two log-rank statistics, normal and log-rank, two-sample Wilcoxon tests, and other linear rank statistics. For two log-rank statistics the correlation estimate of Wei and Lachin (1984) may be used.

In principle, there is no difficulty in estimating correlations, leading to an asymptotically valid use of formula (1). This general strategy for using a global test statistic is now illustrated by two examples.

## 4. Examples

*Example* 1: *A Crossover Trial of Chronic Respiratory Disease*

Seventeen patients with asthma or chronic obstructive airways disease entered a randomized, double-blind crossover trial of an inhaled active drug versus placebo. Each patient received active drug and placebo for consecutive 4-week periods in a random order. The main purpose was to study the drug's possibly harmful effect on lung mucociliary clearance, and analysis of those results produced no evidence of harm. In addition, standard respiratory function measures were taken at the end of both treatment periods. These were peak expiratory flow rate (PEFR), forced expiratory volume ($FEV_1$), and forced vital capacity (FVC), the latter two being expressed as a percentage of the predicted value for that patient's age, sex, and height in the normal population. In this trial the drug or placebo was given in addition to each patient's normal treatment for respiratory disease. A secondary question was whether the addition of this extra inhaled drug could further improve respiratory function.

For each measure there were no signs of period or carryover effects, so that the univariate analysis of drug versus placebo was performed using paired $t$-tests as follows:

| | % predicted $FEV_1$ | % predicted FVC | PEFR |
|---|---|---|---|
| Mean difference (Drug − Placebo) | +7.56% | +4.81% | +2.29 l/min |
| Standard deviation of difference | 18.53% | 10.84% | 8.51 l/min |
| $t$-value | +1.63 | +1.77 | +1.11 |

All three measures showed a mean improvement on active drug but none achieved statistical significance at the 5% level. Thus, Bonferroni correction would lead to a conclusion of no improvement on active drug. Instead we now use formula (3) to assess the collective evidence of drug benefit, taking into account associations between the three measures.

The correlation matrix $\mathbf{S}$ for the three measures' paired differences and its inverse $\mathbf{S}^{-1}$ are as follows:

$$\mathbf{S} = \begin{array}{c} \\ FEV_1 \\ FVC \\ PEFR \end{array} \begin{array}{c} \begin{array}{ccc} FEV_1 & FVC & PEFR \end{array} \\ \begin{bmatrix} 1 & .095 & .219 \\ .095 & 1 & .518 \\ .219 & .518 & 1 \end{bmatrix} \end{array}; \quad \mathbf{S}^{-1} = \begin{array}{c} \begin{array}{ccc} FEV_1 & FVC & PEFR \end{array} \\ \begin{bmatrix} 1.051 & .028 & -.245 \\ .028 & 1.368 & -.715 \\ -.245 & -.715 & 1.424 \end{bmatrix} \end{array}.$$

$\mathbf{J}'\mathbf{S}^{-1}$, the column sums of $\mathbf{S}^{-1}$, are .834, .681, and .464 for $FEV_1$, FVC, and PEFR, respectively. Thus, $FEV_1$ has the greatest weight in the global test statistic, since it is less highly correlated with the other two measures. From formula (3) the global test statistic $t$ equals 2.19. Based on the asymptotic normality of formula (3) and our preliminary simulation studies, one is able to assert that the overall evidence that the extra inhaled drug improved respiratory function is around the 5% level of statistical significance.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

*Example* 2: *A Trial of Two Treatments for Metastatic Colorectal Cancer*
A group sequential trial of two systemic chemotherapy regimens for metastatic colorectal carcinoma, MOF-Strep versus MTX-FU, was designed to enroll (at most) five groups of 17 patients per arm (Geller et al., 1984). The major endpoint was tumour response after 2 months' treatment. However, patient survival time was also considered important. Data for the first group of 17 patients per arm are analyzed here. There were six tumour responses on MOF-Strep and one on MTX-FU, yielding an uncorrected $\chi^2 = 4.50$ for $P = .034$. For survival, the log-rank test yielded $\chi^2 = 2.11$ for $P = .15$, indicating a slight superiority of MOF-Strep. The variance of the log-rank test was 4.44 and there were 18 deaths all together, 7 on MOF-Strep and 11 on MTX-FU. One patient on MOF-Strep responded to treatment and also died. From formula (7), the correlation between the (square roots of the) univariate test statistics was estimated to be .486, i.e., a positive correlation between failure to respond and death. The global test statistic $z$ was calculated using formula (1) to be 2.07, so that $P = .038$.

Suppose this group sequential trial has overall Type I error rate equal to .05 and fixed nominal significance levels, as in Pocock (1977), for the above global test statistic. Then, to stop the trial requires a $z$ test statistic of at least 2.41 and this was not achieved. Thus, even with the two endpoints combined, the results were not extreme enough to stop the trial at the first analysis.

## 5. Discussion

Inevitably, there is no unique, optimal strategy for the use of significance testing when analyzing multiple endpoints in clinical trials. This paper has explored two quite different options, the use of Bonferroni correction and various extensions of a global test statistic proposed by O'Brien (1984). The main advantage of Bonferroni correction is its simplicity, and its slight conservatism is unlikely to be a serious problem. Thus, Bonferroni correction remains useful for avoiding overinterpretation of a set of univariate $P$-values for multiple endpoints. However, since Bonferroni correction utilizes only the most extreme of $k$ $P$-values, it fails to make efficient use of the collective data, particularly in circumstances where one expects several endpoints to behave similarly. One possible alternative is to modify the Bonferroni procedure to take account of more than one $P$-value, as considered by Simes (1986).

The main value of the global test statistic (1) is its applicability to any set of asymptotically normal test statistics, as explored in Section 3. Multivariate methods have often been confined to quantitative data, whereas clinical trials frequently generate binary and censored data. Estimating the correlation matrix $\boldsymbol{\Sigma}$ for asymptotically normal test statistics obtained from such nonnormal data was illustrated by formulae (5) and (7). Another issue is the robustness of the asymptotically normal statistic for finite sample sizes, and this depends primarily on the validity of each univariate normal approximation. O'Brien (1984) has previously reported encouraging findings for quantitative data but we intend to undertake further simulation studies, e.g., for binary and survival data, to explore this issue.

Formula (1) assumes all endpoints are equally important, i.e., the approach is optimal for alternative hypotheses of equal (standardized) magnitude for all endpoints. However, one simple method of attaching unequal priorities (weights) to the various endpoints is as follows. Consider an alternative hypothesis in which $k$ endpoints have standardized treatment differences $\mu/w_1, \ldots, \mu/w_k$. Then the optimal test replaces formula (1) by

$$\frac{\mathbf{J}'(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{W}\mathbf{z}}{[\mathbf{J}'(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W})^{-1}\mathbf{J}]^{1/2}},$$

where $\mathbf{W}$ is a diagonal weighting matrix with elements $w_1, \ldots, w_k$. Both the relative clinical importance of the endpoints and their relative statistical power to detect realistic treatment

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

differences could also play a role in determining such weights. However, one should specify any unequal weights beforehand in order to avoid subjectivity.

Even with endpoints of equal priority, the generalized least squares principle produces a test statistic that is a weighted mean of $k$ standardized normal deviates. Thus, endpoints that are more highly correlated with one another make smaller individual contributions. O'Brien (1984) also considered an alternative procedure based on ordinary least squares that uses the *unweighted* mean of $k$ standardized normal deviates. That is, formula (1) is replaced by $k\bar{z}/(\mathbf{J}'\boldsymbol{\Sigma}^{-1}\mathbf{J})^{1/2}$, which is also a standardized normal deviate under the global null hypothesis. With only two endpoints, the two procedures are identical. For $k > 2$ endpoints, equal correlations among the endpoints are unlikely and therefore the weighted statistic should be more powerful.

However, for certain correlation matrices it is possible for the weighted statistic (1) to have negative weights, which seems untenable from a practical viewpoint. In our experience this appears likely to arise when one is attempting to combine data from diverse endpoints which have an irregular correlation structure. For instance, the crossover trial described in Example 1 had an additional and less widely-used variable, the penetration index (PI), which measures the ability of a deep inhalation to reach small airways. Expanding the analysis to four variables led to the following estimated correlation matrix $\mathbf{S}$ derived from patient paired differences:

$$\begin{bmatrix} 1 & .095 & .219 & -.162 \\ & 1 & .518 & -.059 \\ & & 1 & .513 \\ & & & 1 \end{bmatrix} \begin{matrix} \text{FEV}_1 \\ \text{FVC} \\ \text{PEFR} \\ \text{PI} \end{matrix}$$

This resulted in weights $\mathbf{J}'\mathbf{S}^{-1}$ of 1.38, 1.51, −1.03, and 1.84 for $\text{FEV}_1$, FVC, PEFR, and PI, respectively. The problem arises because PI is correlated with PEFR only, whereas PEFR is also correlated with FVC and $\text{FEV}_1$. The original analysis without PI provided a more logical set of positive weights. Alternatively, one could have resorted to an unweighted test statistic for all four variables.

O'Brien (1984) also considers an alternative nonparametric approach to combining several quantitative endpoints, which will be useful for small data sets where one is unable to rely on asymptotic normality.

One fundamental issue is deciding when a global test statistic is appropriate. Most clinical trials have multiple endpoints, but they are often disparate features of patient response unsuitable for combining. For instance, in primary prevention trials of coronary heart disease it would be inappropriate to combine myocardial infarctions and noncardiovascular deaths into a single global test, since they are totally different outcomes. Global test statistics are more realistic when several endpoints measure closely related aspects of patient response, as in Example 1. Other examples might be in psychiatric illness or rheumatoid arthritis with several measures of symptomatic improvement. Use of a global test statistic may be contentious in some situations. For instance, Example 2 combines tumour response (a short-term measure of drug activity) and patient survival (a more "patient-oriented" assessment of overall benefit). However, since a tumour response usually enhances survival, such a combination may have some merit.

The analysis strategy for multiple endpoints should affect the design. First, awareness of the difficulties in interpreting multiple endpoints should help to avoid an unnecessary excess of endpoints. A clear statement in the trial protocol of the priorities among endpoints, including the possible selection of a single primary endpoint, is desirable. Indeed, the main value of a global test statistic may be in analyzing secondary endpoints, leaving the primary endpoint for univariate analysis. It may sometimes be useful to predefine certain sets of endpoints that assess specific aspects of patient response, each set requiring a separate

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

collective analysis. Also, the method of analysis, Bonferroni correction or a global test statistic, should be specified in the study protocol to avoid any post hoc selection.

In the presentation of results, a global test statistic could either replace or complement the univariate analysis of each endpoint. In order to preserve effective communication with nonstatisticians, the latter approach may be preferable. However, a nonsignificant global test would clearly inhibit any claims of treatment difference for individual endpoints.

This paper has concentrated on significance testing for multiple endpoints, but estimation methods should also be considered. By a procedure analogous to Bonferroni correction, one could widen each univariate confidence interval but this may be unnecessary since uncorrected confidence intervals are usually sufficiently wide to deter exaggerated claims of treatment difference.

Methods for defining multivariate confidence intervals, as in Miller (1981), are difficult to present visually except for the bivariate case. One possibility is to obtain a single confidence interval of treatment difference from the generalized least squares estimation in formula (1). This assumes that the true standardized difference is the same for all endpoints, so this approach may be too abstract for general use.

Other problems worth exploring include the use of multiple endpoints in interim analyses, the extension to more than two treatments as discussed by O'Brien (1984), adjustment for prognostic factors, and methods of assessing the required size of trials with multiple endpoints. Thus, further research is needed to better integrate the concept of a global test statistic into the clinical trial statistician's repertoire, with emphasis placed on the feasibility of implementing such methods in actual trials.

## Acknowledgements

Nancy L. Geller was supported by American Cancer Society Scholar Award #SG-138. We greatly appreciate helpful discussions with Peter Armitage, Derek Cook, Austin Heady, and Max Parmar in the preparation of this paper. We also thank Robert Lee for undertaking the simulation studies and Eva Chan for programming assistance.

## Résumé

Les comparaisons de traitements dans les essais cliniques randomisés mettent, en général, en jeu plusieurs critères finaux, de telle sorte que les tests de signification conventionnels peuvent augmenter sérieusement l'erreur totale de Type I. Une option possible est de sélectionner un seul critère primaire pour faire une inférence statistique correcte, mais ce n'est pas toujours possible. Une autre approche est de faire la correction de Bonferroni (c'est-à-dire multiplier chaque *P*-valeur par le nombre total de critères). Son conservatisme pour des critères corrélés est étudié pour des données suivant une loi normale multidimensionnelle. Une troisième approche consiste à trouver un test statistique global approprié, et cet article étudie un tel test, applicable à tout ensemble de statistiques de tests asymptotiquement distribués suivant une loi normale. Avec cette démarche générale, on peut aussi bien considérer des critères quantitatifs, binaires ou de survie. L'article présente aussi deux exemples, et discute les mérites respectifs des différentes stratégies proposées.

## References

Armitage, P. and Hills, M. (1982). The two-period crossover trial. *Statistician* **31,** 119–131.

Armitage, P. and Parmar, M. (1986). Some approaches to the problem of multiplicity in clinical trials. *Proceedings of the* XIIIth *International Biometric Conference.* Seattle: Biometric Society.

Geller, N. L., Kemeny, N., Yagoda, A., Cheng, E., Sordillo, P., and Hollander, P. (1984). Randomized clinical trial for colorectal carcinoma with planned interim data analysis: A first experience and implications for future design. *Proceedings of the American Association for Cancer Research* **25,** 159.

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

Meier, P. (1975). Statistics and medical experimentation. *Biometrics* **31,** 511–529.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

Miller, R. G. (1981). *Simultaneous Statistical Inference.* New York: Springer-Verlag.
O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40,** 1079–1087.
Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.
Press, S. J. (1972). *Applied Multivariate Analysis.* New York: Holt, Rinehart & Winston.
Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73,** 751–754.
Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79,** 653–661.
Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika* **69,** 297–302.

*Received March* 1986; *revised February* 1987.

## Appendix

In problems of response and survival, the data can be represented as $N$ independent vectors $(R_i, X_i, \Delta_i, Z_i)$, $i = 1, \ldots, N$, where $R$ denotes the response indicator (1 = response, 0 = nonresponse), $X$ denotes the time to failure or censoring, $\Delta$ denotes the failure indicator (1 = failure, 0 = censored), and $Z$ denotes the treatment indicator (1 = treatment A, 0 = treatment B). In what follows all the arguments will be made under the null hypothesis of no treatment effect on response or survival. We shall also condition on the $Z$'s; however, we will assume that as $N$ grows the proportion of treatment A will converge to a constant $\mu$.

The proportions test, given by (4), can also be written as a normalized sum of independent mean zero random variables. If we denote by $T_1$

$$T_1 = \sum_{i=1}^{N} (R_i - \pi)(Z_i - \bar{Z}), \tag{A.1}$$

where $\pi$ denotes the true probability of response and $\bar{Z} = \sum z_i/N$, then (4) is equal to $T_1/\hat{V}_1^{1/2}$, where $\hat{V}_1 = n_1 n_2 \tilde{p}(1 - \tilde{p})/N$.

It will be convenient to express the log-rank test as a stochastic integral of a counting process. That is, the log-rank test given by (6) is also equal to $T_2/\hat{V}_2^{1/2}$, where

$$T_2 = \sum_{i=1}^{N} \int dN_i(u)\{Z_i - \bar{Z}(u)\}, \tag{A.2}$$

where, with $I(A)$ denoting the indicator function of the event $A$,

$$N_i(u) = I(X_i \leqslant u,\ \Delta_i = 1),$$

$$\bar{Z}(u) = \sum_{j=1}^{N} Z_j I(X_j \geqslant u) \Big/ \sum_{j=1}^{N} I(X_j \geqslant u).$$

Using what now has become standard in the theory of counting processes (see Andersen and Gill, 1982; Gill, 1980), we can express the statistic $T_2$ as

$$\sum_{i=1}^{N} \int dM_i(u)\{Z_i - \bar{Z}(u)\}, \tag{A.3}$$

where $dM_i(u) = dN_i(u) - d\Lambda(u)I(X_i \geqslant u)$, and $\Lambda(u)$ denotes the cumulative hazard function of the underlying survival time. If we define the limit of $\bar{Z}(u)$ as $\mu(u)$, then (A.3) can be written as a sum of two terms, namely

$$\sum_{i=1}^{N} \int dM_i(u)\{Z_i - \mu(u)\} \tag{A.4}$$

$$- \sum_{i=1}^{N} \int dM_i(u)\{\bar{Z}(u) - \mu(u)\}. \tag{A.5}$$

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions

Using arguments similar to Tsiatis (1982, §3), we can show that $N^{-1/2}$ times the expression in (A.5) converges in probability to zero. Hence, (A.3) can be approximated by (A.4), a sum of independent mean zero random variables. Therefore, the random vector $(T_1, T_2)$ is asymptotically equivalent to a sum of mean zero random vectors

$$\sum_{i=1}^{N} [(R_i - \pi)(Z_i - \bar{Z}), \int dM_i(u)\{Z_i - \mu(u)\}].$$

Standard results for sums of independent random vectors can be used to show that the joint distribution of $(T_1, T_2)$, properly normalized, will converge to a bivariate normal distribution with mean 0, variance 1, and correlation equal to $\text{cov}(T_1, T_2)/(V_1V_2)^{1/2}$, where

$$\text{cov}(T_1, T_2) = \sum_{i=1}^{N} \text{E}[(R_i - \pi)(Z_i - \bar{Z}) \int dM_i(u)\{Z_i - \mu(u)\}].$$

In general, $\text{cov}(T_1, T_2)$ may be estimated by substituting $\bar{p}$ for $\pi$, $\bar{Z}(u)$ for $\mu(u)$, and the Nelson estimate of the cumulative hazard function for $\Lambda$. A simplification of the estimate can be obtained in the special case that the underlying censoring distribution is independent of treatment, as is the case in most randomized clinical trials. In such instances, under the null hypothesis of no treatment effect on response or survival, the random vector $(R, X, \Delta)$ is independent of $Z$. Therefore, $(R_i, X_i, \Delta_i)$, $i = 1, \ldots, N$, are identically and independently distributed. It is also clear that $\mu(u)$, the proportion expected on treatment A among individuals at risk at time $u$ will be independent of time, i.e., $\mu(u) = \mu$ for all $u$. Therefore, $\text{cov}(T_1, T_2)$ is equal to

$$\sum (Z_i - \bar{Z})(Z_i - \mu)\text{E}\{(R_i - \pi) \int dM_i(u)\}.$$

However, $\text{E}\{(R_i - \pi) \int dM_i(u)\}$ is the same for each $i$ and therefore can be estimated by

$$\sum [(R_i - p) \int \{dN_i(u) - dN(u)I(X_i \geq u)/Y(u)\}]/N. \tag{A.6}$$

We note that $\sum p \int \{dN_i(u) - dN(u)I(X_i \geq u)/Y(u)\} = 0$, and therefore (A.7) is equal to

$$N^{-1}\left[\sum R_i \int dN_i(u) - \int dN(u) \sum_{i=1}^{N} \{I(X_i \geq u, R_i = 1)\}/Y(u)\right].$$

Since $\int dN_i(u) = \Delta_i$, then the sum $\sum R_i \int dN_i(u)$ corresponds to the number of individuals who both respond and die, which will be denoted by $R_d$. We shall also define $\bar{p}(u) = \sum I(X_i \geq u, R_i = 1)/Y(u)$, which is the proportion of individuals who respond among those at risk at time $u$. Therefore, if we denote by $t_1, \ldots, t_k$ the distinct death times, then (A.6) is equal to

$$N^{-1}\{R_d - \sum_{j=1}^{k} \bar{p}(t_j)\}.$$

Consequently, an estimate for $\text{cov}(T_1, T_2)$ under the assumption of equal censoring would be

$$N^{-1} \sum_{i=1}^{N} (Z_i - \bar{Z})^2\{R_d - \sum_{j=1}^{k} \bar{p}(t_j)\}.$$

Noting that $\sum_{i=1}^{N} (Z_i - \bar{Z})^2 = n_1n_2/N$, we can write the estimate for the correlation coefficient as

$$\frac{\{R_d - \sum_{j=1}^{k} \bar{p}(t_j)\}}{N[NV_2\bar{p}(1 - \bar{p})/n_1n_2]^{1/2}}.$$

## Appendix References

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large-sample study. *Annals of Statistics* **10,** 1100–1120.

Gill, R. D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77,** 855–861.

This content downloaded from 132.174.255.223 on Mon, 5 Aug 2013 17:51:05 PM
All use subject to JSTOR Terms and Conditions