Payne, J. N., Coy, J., Milner, P. C. and Patterson, S. (1993) Are deprivation indicators a proxy for morbidity?: a comparison of the prevalence of arthritis, depression, dyspepsia, obesity and respiratory symptoms with unemployment rates and Jarman scores. *J. Publ. Hlth Med.*, **15**, 161–170.
Pocock, S. J., Shaper, A. G., Cook, D. G., Phillips, A. N. and Walker, M. (1987) Social class differences in ischaemic heart disease in British men. *Lancet*, ii, 197–201.
Power, C., Manor, O. and Fox, A. J. (1991) *Health and Class: the Early Years*. London: Chapman and Hall.
Rose, G. A. (1962) The diagnosis of ischaemic heart pain and intermittent claudication in field surveys. *WHO Bull.*, **27**, 645–658.
Royal College of General Practitioners, Office of Population Censuses and Surveys, and Department of Health and Social Security (1986) *Morbidity Statistics from General Practice 1981–82; Third National Survey*. London: Her Majesty's Stationery Office.
——— (1990) *Morbidity Statistics from General Practice 1981–82. Third National Study: Socio-economic Analyses*. London: Her Majesty's Stationery Office.
Samphier, M. L., Robertson, C. and Bloor, M. J. (1988) A possible artefactual component in specific cause mortality gradients. *J. Epidem. Commty Hlth*, **42**, 138–143.
Scambler, A., Scambler, G. and Craig, D. (1981) Kinship and friendship networks and women's demand for primary care. *J. R. Coll. Gen. Pract.*, **31**, 746–750.
Skrimshire, A. (1978) Area disadvantage, social class and the health service. *Report*. Social Evaluation Unit, Department of Social and Administrative Studies, University of Oxford, Oxford.
Stansfield, S. A. and Marmot, M. G. (1992) Social class and minor psychiatric disorder in British civil servants: a validated screening survey using the General Health Questionnaire. *Psychol. Med.*, **22**, 739–749.
Sterling, P. and Eyer, J. (1981) Biological basis of stress-related mortality. *Socl Sci. Med.*, **15**, 3–42.
Szreter, S. (1988) The importance of social intervention in Britain's mortality decline 1859–1914. *Socl Hist. Med.*, **1**, 1–37.
Tudor Hart, J. (1971) The inverse care law. *Lancet*, i, 405–412.
Wadsworth, M. E. J. (1991) *The Imprint of Time: Childhood, History and Adult Life*. Oxford: Clarendon.
Wadsworth, M. E. J., Butterfield, W. and Blaney, R. (1971) *Health and Sickness: the Choice of Treatment*. London: Tavistock.
Williams, G. H. (1989) Hope for the humblest? The role of self-help in chronic illness: the case of ankylosing spondylitis. *Sociol. Hlth Ill.*, **11**, 135–159.
Zola, I. (1973) Pathways to the doctor: from person to patient. *Socl Sci. Med.*, 7, 677–689.

J. R. Statist. Soc. A (1996)
159, *Part* 1, *pp.* 93–110

Multiplicity Considerations in the Design and Analysis of Clinical Trials

By RICHARD J. COOK† and VERN T. FAREWELL

University of Waterloo, Canada

[Received December 1994. Revised May 1995]

SUMMARY

The need for efficient use of available resources in medical research has led to the increased appeal of clinical trial designs based on multiple responses, multiple treatment arms and repeated tests of significance. In recent years there has been considerable methodological work pertaining to these types of multiple comparison, with the common objective typically being the control of the experimental type I error rate. Here we reconsider the appropriateness of these objectives in a variety of contexts and suggest that multiple-comparison procedures are frequently adopted unnecessarily. In particular we argue that, provided that a select number of important well-defined clinical questions are specified at the design, there are situations in which multiple tests of significance can be performed without control of the experimental type I error rate. The primary restriction for this to be reasonable is that test results are interpreted marginally.

Keywords: MULTIPLE OUTCOMES; MULTIPLE TREATMENT ARMS; POWER; *p*-VALUE; SEQUENTIAL TESTS; SIZE; TYPE I ERROR RATE

1. INTRODUCTION

The need for comprehensive and efficient methods for evaluating therapeutic interventions has led to the increased use of more complex experimental designs in clinical research in recent years. In particular, clinical trials are now frequently designed based on multiple responses (O'Brien, 1984; Pocock *et al.*, 1987; Wei *et al.*, 1989), multiple treatment arms (Mau, 1988; Liu *et al.*, 1993), repeated tests of significance (Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983) and various combinations of these features (O'Brien, 1984; Tang *et al.*, 1989; Lin, 1991; Proschan *et al.*, 1994; Hughes, 1993; Cook and Farewell, 1995; Cook, 1994, 1996). The trend for the use of these more complex designs has led to an increased interest in issues surrounding multiplicity in clinical trials.

It is most common to address multiplicity concerns by attempting to formulate statistical tests of significance which control an experimental type I error rate at some specified level. The purpose of this paper is to consider whether such an approach is the most natural in certain contexts of clinical trial research. In particular, consideration is given to scenarios in which multiple tests of significance might be performed without direct control over some overall type I error rate. In addition, guidelines for multiplicity adjustments, when appropriate, are examined. Central to the discussion will be the need to specify well-defined clinical questions and the role that they will play in formulating treatment recommendations.

†*Address for correspondence*: Department of Statistics and Actuarial Science, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada.
E-mail: rjcook@jeeves.uwaterloo.ca

 0035–9238/96/159093

In the next section alternative roles and views of tests of significance are reviewed. In Section 3, the rationale for the use of multiple outcomes is discussed, several methods proposed for trials with multiple outcomes are considered, and the motivation for the particular multiplicity adjustments is re-examined. Clinical trials with multiple treatment arms are examined in Section 4, whereas sequential procedures based on repeated tests of significance are described in Section 5. Some general concluding remarks are made in Section 6.

2. HYPOTHESIS TESTING IN CLINICAL TRIALS

2.1. *Background*

Consider a phase III clinical trial with the objective of comparing the efficacy of an experimental therapy (treatment 1) with a standard therapy (treatment 2) on the basis of a univariate response of interest. Although response variables with a wide variety of distributions are used in clinical trials, it is sufficient for the present purposes to focus on normally distributed responses. Thus, suppose that $2n$ patients are serially accrued and randomized in pairs to treatment groups with $Y_{ij} \sim N(\theta_i, \sigma^2)$ representing the response variate corresponding to the jth patient randomized to the ith treatment, $j = 1, \ldots, n$, $i = 1, 2$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in})'$, $\bar{Y}_i = \Sigma_{j=1}^{n} Y_{ij}/n$, $i = 1, 2$, and $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$. If $\delta = \theta_1 - \theta_2$ is the parameter of interest, then $D(\mathbf{Y}, \delta) = \sqrt{n}(\bar{Y}_1 - \bar{Y}_2 - \delta)/\sqrt{2}\sigma$ is a standard normal pivotal quantity on which inference for δ can be based. In particular, if H_0: $\delta = \delta_0$ *versus* H_A: $\delta \neq \delta_0$ are the hypotheses of interest, $D(\mathbf{Y}, \delta_0)$ is the relevant standard normal discrepancy measure under the null hypothesis. Since $\delta_0 = 0$ will be of interest most often, it will be assumed subsequently, with $D(\mathbf{Y})$ used to denote the corresponding discrepancy measure, $D(\mathbf{Y}, \delta = 0)$.

Given observations $\mathbf{y}$, let $d(\mathbf{y})$ denote the realization of $D(\mathbf{Y})$. A significance level, or p-value, associated with a test of the null hypothesis H_0: $\delta = 0$, is then defined as

$$p = \text{pr}(|D(\mathbf{Y})| > |d(\mathbf{y})|; \delta = 0).$$

The significance level, as discussed by Fisher (1971), is intended to be used by scientists to assess the plausibility of the null hypothesis given the data. More specifically, 'moderate' p-values reflect little evidence against the null hypothesis, whereas 'small' p-values indicate that either

(a) a 'rare' event has occurred or
(b) there is evidence that the null hypothesis is not true;

typically the latter explanation for small p-values is adopted. The question then remains about what constitutes a small p-value.

Within Fisher's framework for signifcance testing, we can conceptualize a generic threshold significance level α such that, if $p < \alpha$, we could reasonably claim that there is sufficient evidence to reject H_0: $\delta = 0$ in favour of the alternative. Corresponding to this α is a critical value $z_{\alpha/2}$, such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ where $\Phi(\ )$ is the cumulative distribution function for the standard normal distribution. Attention is often directed towards $z_{\alpha/2}$ so that information regarding H_0 is quantified on the D-scale (e.g., if $|d(\mathbf{y})| > z_{\alpha/2}$, H_0: $\delta = 0$ should be rejected). For historical, computational and other practical reasons, Fisher (1946) provided tables of the standard

normal, χ^2-, t- and F-distributions which indicated selected percentage points of these distributions. This led to the widespread adoption of the 5% and 1% threshold significance levels in tests of hypotheses. In particular, $p < 0.05$ is often thought of as providing sufficient evidence against H_0 to reject it, whereas $p < 0.01$ provides strong evidence against it. Although Fisher embraced the concept of evidence which would lead to the rejection of a null hypothesis, he opposed, at times vehemently, the view of significance testing solely as a decision-making process and the interpretation of α as a rejection rate (Fisher, 1955, 1960). Rather, he stressed the value of the *strength* of evidence, reflected by the actual p-value, and emphasized the role that it could play in subsequent inferences (Fisher, 1973). Cox (1977) presented a similar view in describing the p-value as a measure of the degree of consistency with the null hypothesis.

The decision theoretic view was adopted by Neyman and Pearson (1928, 1933), however, leading to the notions of type I and type II error rates, or equivalently the size and power, of statistical tests of hypotheses. Specifically, Neyman and Pearson (1933) would interpret Fisher's threshold significance level α as the type I error rate of a test; the probability of rejecting the null hypothesis if it is correct. A more formal interpretation relies on the notion of a hypothetically infinite population of experiments corresponding to repetitions of the performed experiment. In this context, the type I error rate of the test can be defined as the percentage of this infinite population of trials leading to erroneous rejection of the null hypothesis. Conversely, a type II error is said to be committed in a single trial if the null hypothesis is not rejected when it is incorrect. The type II error rate, denoted β, is defined as the frequency with which this error would occur in the aforementioned infinite population of trials with δ fixed at a value of interest, denoted $\delta = \delta_A$ ($\delta_A \neq 0$). The power of a particular study to detect an effect δ_A is the rate at which the null hypothesis would be rejected and is denoted by $1 - \beta$. The concept of power is of greatest relevance at the design stage of clinical investigations, proving very useful in the specification of sample size requirements of randomized trials.

Although Fisher did not subscribe to the notion of error rates, and therefore did not discuss power and its relationship to sample size calculations *per se*, he did refer to the notion of 'increasing sensitivity' (Fisher (1971), p. 22). In the context of the tea tasting experiment, for example, he stated that

> 'without specific precautions, even a definite sensory discrimination would have little chance of scoring a significant result'.

In the subsequent paragraph, he indicates that sensitivity to departure from the null hypothesis 'can usually be achieved . . . by increasing the size of the experiment'. Thus, although not formally recognized as such, the notion of planning the sample size of an experiment with a view to improving the ability to detect departures from the null hypothesis is consistent with Fisher's views.

Interestingly, despite their different views on the roles of tests of significance, it is likely that both Fisher and Neyman would be very uncomfortable with the widespread adoption of the 5% threshold significance level in experimental research today. The threshold significance level, or type I error rate as viewed by Neyman and Pearson, was originally viewed as a flexible criterion that should be determined in conjunction with the type II error rate, whereas Fisher, particularly later in life, strongly objected to the adoption of such a narrow view of significance testing. An

excellent overview of the long-standing debate between Fisher and Neyman is given in Lehman (1993). As pointed out by Lehman (1993), the centre of this controversy appears to involve their differing views on the purpose of hypothesis tests. Fisher viewed the objective as one of making '*inductive inferences*' (e.g. drawing conclusions), whereas the Neyman–Pearson framework is more closely tied to '*inductive behaviour*' (e.g. making decisions, and presumably acting on them) which involves either rejecting or failing to reject (initially termed 'accepting' by Neyman and Pearson (1928)) the null hypothesis (Fisher, 1955). Thus, for example, Fisher's concern with quantifying the strength of evidence against the null hypothesis is not much reflected in the Neyman–Pearson framework. As pointed out by Cox (1977), the contrast between significance tests as an aid in summarizing evidence and decision procedures is implicit or explicit in much of the discussion of significance tests.

Although we favour the interpretation of tests of significance as a means of assessing strength of evidence, much of the discussion that follows is phrased in the Neyman–Pearson terminology for pragmatic and semantic reasons. The Neyman–Pearson structure provides a simple and convenient framework within which clinical trials may be designed and discussed. This artificial simplicity to this framework is analogous to those used for other aspects of trial design. For example, two-sample tests are often used as a basis for design even if the planned analyses involve multiple regression, and constant hazard rates are often assumed for planning trials with time-to-event outcomes. Although this structure is convenient for design purposes, the analysis should retain an emphasis on the strength of evidence. Thus, we support Fisher's emphasis on the strength of evidence and use this to rationalize the appeal of marginal inferences. In the Neyman–Pearson context this in turn implies interest in marginal error rates.

2.2. *Multiplicity Adjustments*

Issues regarding multiple tests of significance primarily arise within the Neyman–Pearson framework with the fundamental concern stated as follows. If a number of significance tests are performed, all with a $100\alpha'\%$ type I error rate, the percentage of experiments with *one or more* false positive errors in the hypothetically infinite population of such experiments may be much larger than $100\alpha'\%$. There are many references on methodology designed to control the type I error rate in fixed sample size experiments with multiple tests of hypotheses (e.g. Hoppe (1993)). The objective of most of these procedures is to develop testing algorithms such that, in the hypothetical population of trials, one or more false positive decisions would be made at most $100\alpha\%$ of the time, where α is specified in advance. For example, if K response variables are of interest in a clinical trial and a test of significance is performed based on each, these multiplicity corrections are designed to provide a method of guarding against an excessive number of false positive decisions in the set of all hypothesis tests.

Cox (1965) has pointed out that probabilities regarding the simultaneous correctness of many statements may not always be of direct relevance, particularly when there is interest in a specific response. This suggests that, in clinical trial designs formally based on two or more responses, multiplicity adjustments may not be necessary if marginal, or separate, test results are *interpreted marginally* and have

implications in *different aspects* of the prescription of the treatments (i.e. response-specific effects are of interest and separate statements regarding them are desired). Thus there may be contexts in which multiple tests of significance should be performed with reference to marginal rather than experimental error rates. If hypothesis tests are primarily directed at marginal inferences, it is then reasonable to specify a maximum tolerable error rate for each specific hypothesis test. This point can be highlighted by careful consideration of the precise nature of the possible type I errors and the meaningfulness of an 'overall' type I error, i.e. at the design stage of a clinical trial an investigator will have to consider the nature and relative frequency of the type I errors that will be tolerated. These ideas are discussed further in Sections 3–5.

Fisher's view of significance tests as a means of inductive inference is consistent with the idea of performing unadjusted marginal analyses. In particular, with several tests directed at comparing two treatments based on several outcomes, the marginal p-values indicate the strength of the evidence against each null hypothesis. Inductive behaviour should be influenced by the information that is available from all the outcome variables, i.e. the evaluation of the *experimental treatment* will be based on all the marginal results. This may involve a subtle balancing of the different types of information, a balancing not easily specified in advance.

In the next sections, specific forms of multiple tests will be discussed to expand on these ideas further.

3. MULTIPLE RESPONSES

3.1. *Medical Motivations for Multiple Outcomes*

For some medical conditions, the response to treatment may be sensibly characterized by a single response variate. In such a situation, a demonstration of treatment efficacy with respect to this response leads directly to treatment recommendations. For example, a live-birth is the natural outcome variable for a study related to the prevention of miscarriages in pregnancy (e.g. Diddle *et al.* (1953)).

In other cases, however, it is difficult to identify a single primary outcome that adequately characterizes response to treatment. For example, O'Brien (1984) described a diabetes study in which 34 related response variables were recorded to characterize the treatment effect on nerve function. It appears that there was no *a priori* reason to examine this constellation of variables in any single specific manner. In arthritis studies, disease activity is reflected in a variety of measured variables including pain, active joint counts, strength and mobility (e.g. Fitzpatrick *et al.* (1987)). Although several variables are usually examined, the question of how to measure the effect of treatments on disease activity remains very ill defined.

Quite a different context in which multiple outcomes arise is when there is genuine interest in treatment effect on two or more important clinical outcomes. In these situations, response variables are linked to specific clinical outcomes. One example of such a study is a stroke prevention trial (North American Symptomatic Carotid Endarterectomy Trial Collaborators, 1991). This study evaluated the effect of carotid endarterectomy *versus* aspirin therapy in reducing the risk of a variety of stroke types characterized in terms of severity and location. The potential surgical benefit on each individual stroke type was of clinical interest. Another example is the clinical trial of co-trimoxazole for cyclospora infections reported by Hoge *et al.* (1995). Patients

presenting with gastrointestinal complaints and showing evidence of cyclospora infection were randomized to receive co-trimoxazole or placebo treatment for 7 days. The outcomes of interest were improvement in clinical symptoms measured on a three-point ordinal scale, a binary indicator of clearance of cyclospora, and relapse. A third example is a kidney transplant study (Friend *et al.*, 1994) in which a 10-day regimen of Campath, to be added to cyclosporin, was compared with a placebo for the prevention of graft rejection. An immunological effect was expected to be demonstrated by results on graft survival but there was separate concern about toxicity which would probably be reflected through infections. All three of these trials are different from the diabetes and arthritis studies in that there is no single clinical concept or construct which is of interest and which could perhaps be defined in terms of the multiple-response variables.

These two scenarios are not mutually exclusive and individual studies may exhibit multiplicity features of both types.

3.2. *Some Current Approaches*

3.2.1. *Notation*

To generalize the notation of Section 2, let $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijK})'$ denote a $K \times 1$ response vector corresponding to the jth patient randomized to the ith treatment, $j = 1, \ldots, n$, $i = 1, 2$. We assume that $\mathbf{Y}_{ij} \sim \text{MVN}(\boldsymbol{\theta}_i, \Sigma)$ where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})'$ and Σ is a known $K \times K$ covariance matrix with diagonal entries denoted $\{\sigma_k^2\}$ and off-diagonal entries denoted $\{\sigma_{kk'}\}$. Let $\mathbf{Y}_{ik} = (Y_{i1k}, \ldots, Y_{ink})'$, $\bar{Y}_{i\cdot k} = \Sigma_{j=1}^{n} Y_{ijk}/n$, $i = 1, 2$, and $\mathbf{Y}_k = (\mathbf{Y}_{1k}', \mathbf{Y}_{2k}')'$, $k = 1, \ldots, K$. This structure leads to the construction of K marginal pivotal quantities $D_k(\mathbf{Y}_k, \delta_k) = \sqrt{n}(\bar{Y}_{1\cdot k} - \bar{Y}_{2\cdot k} - \delta_k)/\sqrt{2}\sigma_k$ which can be used for inference on $\delta_k = \theta_{1k} - \theta_{2k}$, the treatment effect for the kth response, $k = 1, \ldots, K$. When $D_k(\mathbf{Y}_k, \delta_k)$ is evaluated at $\delta_k = 0$, we denote it by $D_k(\mathbf{Y}_k)$.

Three testing strategies for such contexts are based on global test statistics, summary measures and marginal testing procedures. The choice of strategy is linked to the reason for adopting multiple outcomes, the nature of the null hypothesis, and by the treatment effect of interest. We do not attempt to review the large literature pertaining to methods for multiple outcomes in clinical research, but we discuss some general frameworks and specific procedures for illustration.

3.2.2. *Global tests*

One hypothesis of frequent interest is of the form H_0: $\boldsymbol{\delta} = \mathbf{0}$ *versus* H_A: $\boldsymbol{\delta} \neq \mathbf{0}$. Although a natural and appropriate analysis would be based on Hotelling's T^2-statistic in many contexts, it is not appropriate for use in clinical trials. One reason relates to the notion of an effect of clinical interest which we shall denote here by $\boldsymbol{\delta}_A = (\delta_{1A}, \ldots, \delta_{KA})'$. A global test involving multiple outcomes would be most natural if the treatment effect is expected to be comparable across responses. For example, we might define the specific effect sizes of clinical interest by $\delta_{kA} = \gamma\sigma_k$, $k = 1, \ldots, K$. Hotelling's T^2-test is sensitive to this type of departure from the null hypothesis, but since it makes no distinction between variables that respond favourably and unfavourably to treatment it is equally sensitive to departures of a similar magnitude that do not lead to clear treatment recommendations.

O'Brien (1984) recognized this difficulty and developed a generalized least squares approach that leads to the 'most powerful' test for detecting treatment effects of the form $\delta_{kA} = \gamma\sigma_k$, $k = 1, \ldots, K$. Variables are assumed to be defined so that shifts in the same direction represent a favourable response to treatment and inference is based on a linear combination of the marginal discrepancy measures. O'Brien's (1984) method will lead to a straightforward treatment recommendation if the null hypothesis is rejected but there may be other departures from the null hypothesis, the detection of which would also be clinically informative. O'Brien sensibly described his procedure as a supplement to univariate methods because it, like Hotelling's T^2-test, does not explicitly indicate the nature of apparent treatment differences. In evaluating therapeutic interventions it is clearly important to identify which responses were affected by the intervention, and what the nature of the effects was.

3.2.3. *Summary measures*

Another approach for testing hypotheses of the form H_0: $\boldsymbol{\delta} = \mathbf{0}$ *versus* H_A: $\boldsymbol{\delta} \neq \mathbf{0}$ is the adoption of a single summary response, at the individual level, based on combining univariate responses in some clinically sensible manner. For example, this is often used in quality-of-life studies when responses are combined (in an unweighted or weighted manner) across various dimensions. For more general applications, when no specific numerical means of combining results is available, O'Brien (1984) proposed the following procedure based on ranks.

The analysis involves ranking all observations across treatment groups and patients, for each response variable. The sum of the ranks assigned to an individual is a summary measure at the patient level reflecting the patient's 'overall' response to treatment relative to all other patients. In the case of parallel group designs of moderate size, two-sample t-tests can be performed based on the summary outcome measure. For smaller trials, inference can be based on the permutation distribution of a relevant discrepancy measure. This method makes general statements about the overall relative performance of the experimental treatment possible. Significant test results provide evidence that one treatment is 'uniformly' better than another, facilitating general clinical recommendations. It has the limitation, however, that since the procedure is based on ranks treatments that have dramatic effects on one or two responses may be undetected if a large number of the other responses reflect neglible benefit or harm.

3.2.4. *Marginal testing procedures*

In some situations, the set of hypotheses of the form H_{k0}: $\delta_k = 0$ *versus* H_{kA}: $\delta_k \neq 0$, $k = 1, \ldots, K$, is of interest. As before marginal discrepancy measures can be defined. Then $D_k(\mathbf{Y}_k) \sim N(0, 1)$ if H_{k0}: $\delta_k = 0$ is true, and $D_k(\mathbf{Y}_k) \sim N(\delta_{kA}\sqrt{n}/\sigma_k\sqrt{2}, 1)$ if $\delta_k = \delta_{kA}$, $k = 1, \ldots, K$. Typically if tests of the marginal hypotheses are performed separately they are done so with a view to controlling the experimental type I error rate.

Bonferroni adjustments provide a simple method for ensuring that the experimental type I error rate is not exceeded. With this approach, marginal analyses are performed based on $d_k(\mathbf{y}_k)$, $k = 1, \ldots, K$. If the experimental type I error rate is α, the marginal nominal type I error rates are taken as α/K. Then H_{k0} is rejected if and only if $|d_k(\mathbf{y}_k)| > z_{\alpha/2k}$. The conservative nature of this procedure is examined by

Pocock *et al.* (1987). Less conservative approaches with similar objectives have been described by Gupta (1963), Simes (1986) and others. Stepwise Bonferroni procedures have been developed by Holm (1979), and more recently Hochberg (1988).

3.3. *Comments*

The two scenarios discussed in Section 3.1 motivating the use of multiple responses can now be considered in more detail.

Consider first the diabetes study described by O'Brien (1984) in which 34 electromyographic variables are recorded to reflect the nerve function response in two groups of patients, one receiving the experimental therapy, the other a standard treatment. The motivation for the use of the large number of variables was the inability to characterize the disease state more specifically. In a scenario such as this, the null hypothesis can be broadly stated as

H_0: there is no improvement in nerve function under the experimental treatment;

the alternative may be one sided and could be stated as

H_A: there is improvement in nerve function under the experimental treatment.

By the very nature of these hypotheses, it is clear that the requirements for a superior treatment are not well defined. As a result, we cannot expect to perform a test of hypothesis, to make statistical inferences and to obtain clear unambiguous clinical recommendations regarding the appropriate treatment and the nature of the expected response. Given this limitation, it is natural to rely on a summary response, such as the rank sum described by O'Brien (1984), to perform tests based on this response and to make broad statements regarding the performance of the experimental treatment.

The adoption of such an approach should be viewed as an acceptable approach only if a small number of clinically important responses cannot be clearly identified. Although a comparative study based on a summary response is possible, it will necessarily retain a strong exploratory flavour. As well, this approach which involves high dimensional multivariate response vectors does not lend itself to trial planning. Clinically important effect sizes are difficult to conceptualize and identify. Hence traditional power and sample size calculations are problematical.

Consider the second scenario in which a smaller number of response variables identified at the design stage of the trial are of joint interest. For example consider the colorectal cancer trial reported by Pocock *et al.* (1987) in which two systemic chemotherapy regimens were evaluated on the basis of tumour response after two months of treatment, and survival time. Pocock *et al.* (1987) performed an analysis based on an extension of O'Brien's (1984) generalized least squares global test statistic. The univariate uncorrected χ^2-statistic for tumour response yields $\chi^2_{(1)} = 4.50$ ($p = 0.034$), whereas the log-rank test for the equality of the survival time distributions leads to $\chi^2_{(1)} = 2.11$ ($p = 0.15$). The combination of these statistics leads to a global statistic with a z-statistic of 2.07 ($p = 0.038$). It has been our experience, and it is not surprising, that physicians find results expressed in terms of global test statistics unintuitive and clinically uninformative. In this case, the value $p = 0.038$ reflects evidence against H_0, but this value could result from dramatic treatment effects based on one response and moderate treatment effects on another, or two

consistently moderate treatment effects. In the former case, interest may lie in which response the large treatment effect is manifested leading to separate analyses to control the experimental type I error rate. Although it was not the intention of Pocock *et al.* (1987) to make clinical recommendations regarding these chemotherapy regimens, no summary statement is provided to facilitate interpretation of the global test result in this example. This may be a natural consequence of adopting such global test statistics for inference.

An alternative analysis strategy for this scenario is to apply Bonferroni, or other types of, multiplicity corrections to univariate analyses. This approach would force us to interpret the test results as providing no evidence of treatment benefit in terms of tumour response or survival. In not identifying the treatment as effective, this analysis does, however, provide clear unambiguous answers to specific clinical questions of interest.

This trial, however, is an example of a study in which global test statistics or multiplicity adjustments are unnecessary. Under the assumption that the two responses were identified as of interest at the design stage of the study, the rationale involves the examination of type I errors. Suppose that two treatments are to be compared, there is a total of K responses with hypotheses of the form H_{k0}: $\delta_k = 0$ *versus* H_{kA}: $\delta_k \neq 0$, $k = 1, \ldots, K$, and K univariate tests are planned, one for each hypothesis. Suppose that the experimental treatment would be recommended for use if H_{k0}: $\delta_k = 0$ was rejected in favour of $\delta_k = \delta_{kA} > 0$ for at least one response. Then a type I error is naturally defined as the identification of a treatment as 'beneficial' and recommending it for use (the behavioural component), when in fact it is of no additional therapeutic value over the standard treatment (e.g. H_{k0} is true for all k, $k = 1, \ldots, K$). Under these circumstances within the Neyman–Pearson framework, it is usually argued that the experimental type I error rate must be controlled to ensure that ineffective treatments are infrequently (with the specific frequency reflected by α) recommended for use. The clinical trial, in this case, represents a decision-making procedure.

It is difficult to envisage, however, that for this chemotherapy clinical trial, and indeed for many similar clinical trials, this analysis and behaviour strategy would be adopted. It is difficult to link the decision-making process regarding the clinical value of the experimental treatment with outcomes of the marginal hypothesis tests in such a simplistic manner. For example, the clinical implications of a trial providing very strong evidence of treatment benefit in terms of tumour response and weak evidence of benefit in terms of survival are very different from those of a trial exhibiting weak evidence of benefit in terms of tumour response and very strong evidence of benefit in terms of survival. Thus the two responses seem to be of interest in their own right and the effect of treatment on both is relevant to patient care. The notion of an experimental error rate is of questionable relevance, particularly in the context of multiple two-sided tests of this sort. If multiple univariate test results have implications on specific responses, then a marginal interpretation is reasonable.

In this colorectal cancer trial, then, it is reasonable to provide answers to the hypotheses regarding tumour response and survival separately. Thus two marginal tests of significance would be performed. These, and of course associated estimation procedures, would be used to evaluate the strength of the evidence for a treatment effect on the two outcomes. As indicated above, it could be very useful for future treatment choices and development to know that there is strong evidence for a

survival benefit but little for an effect on tumour response, a result which might have been viewed as unlikely at the time of trial design. With these considerations in mind, it is difficult to view this procedure as a decision-making process which lends itself to a well-defined experimental type I error.

A further justification for marginal inferences follows. Suppose that some commonality of the treatment effect was expected and an analysis was to be based on O'Brien's (1984) generalized least squares statistic. This requires the marginal component test statistics to be calculated. If at this stage strong evidence of a treatment benefit based on a particular response was observed and evidence of harm based on another, the univariate analyses might be relied on to draw appropriate marginal inferences. This reasonable possibility supports the view that marginal analyses are fundamentally of primary interest.

In some trials, multiple outcomes may be of interest in fundamentally different manners. There may be equivalence questions concerning some outcomes and increased efficacy or reduced toxicity questions concerning others. This highlights another advantage of performing several univariate analyses in that design requirements are more easily characterized. In particular, suppose that H_{k0}: $\delta_k = 0$ *versus* H_{kA}: $\delta_k \neq 0$ is to be tested with a marginal size α_k-test. If a power requirement of $1 - \beta_k$ is identified at a clinically important effect size $\delta_k = \delta_{kA}$, then we require $n_k \geqslant 2\sigma_k^2(z_{\alpha_k/2} + z_{\beta_k})^2/\delta_{kA}^2$. If it is desirable to demonstrate equivalence based on the lth response, suppose that γ_l defines a point of therapeutic equivalence and $(\gamma_{1l}, \gamma_{2l})$ defines a region of therapeutic equivalence. Following the arguments of Fleming (1990), to ensure identification of inferior, equivalent and superior experimental treatments, we require $n_l \geqslant 4\sigma_l^2 z_{\alpha_l/2}^2/(\gamma_{1l} - \gamma_{2l})^2$. Whatever the purpose of the marginal analyses, choosing a final sample size of $n = \max(n_1, n_2, \ldots, n_K)$ ensures that all marginal power and equivalence requirements are satisfied.

A referee has pointed out that the analysis of different aspects of the effect of treatment frees the clinical investigator from the need to specify a single primary outcome variable. We agree with this and would state further that this freedom does not support the examination of endless outcome variables in an effort to find one generating statistical significance, even if they are associated with *post hoc* rationalizations of their importance. Thought and discipline are required at the design stage of the trial to define outcomes of interest.

4. MULTIPLE TREATMENT ARMS

4.1. *Medical Motivation and Current Approaches*

Although the simplicity of the two-armed randomized trial is attractive, the complexity of modern therapies means that increasingly clinical trials are designed with multiple treatment arms. These arms might involve increasing dosages of an experimental treatment, cumulative combination therapies or simply multiple different treatments. Such designs are often viewed as an efficient means for simultaneously assessing the efficacies of a variety of treatment options. For example, several active treatments can be compared with one another and/or with a common control group receiving standard medical care.

We return to the notation of Section 2.1 and generalize it as follows. Suppose that patients are serially accrued and randomized to one of I treatment groups in blocks of size I with $Y_{ij} \sim N(\theta_i, \sigma^2)$ the response variate for the jth patient randomized to the

ith treatment group, $j = 1, \ldots, n$, $i = 1, \ldots, I$. When different treatments or multidrug therapies are used in a clinical trial, the most relevant null and alternative hypotheses are not always obvious. Frequently global tests are proposed to assess whether or not there is evidence of differences in the efficacy across all treatments. In these contexts hypotheses take the form H_0: $\theta_1 = \theta_2 = \ldots = \theta_I$ *versus* H_A: $\theta_i \neq \theta_{i^*}$ for at least one $i \neq i^*$, $i, i^* = 1, \ldots, I$. These tests might be based on F-statistics arising from analysis-of-variance tables or global tests regarding coefficients in a regression model. Such procedures are typically insufficient by themselves, however, as rejection of this null hypothesis does not indicate which treatment arms provided evidence against H_0; this is analogous to the criticism levelled against the use of Hotelling's T^2-test in the context of trials with multiple outcomes. Therefore, on rejection of such stringent null hypotheses, inference regarding specific contrasts is desirable to help to identify and characterize the effect detected.

One approach, often used in the analysis of variance for factors with I levels, is the use of $I - 1$ orthogonal contrasts effectively to partition the treatment sum of squares and hence to ensure that the component χ^2 discrepancy measures are independent. Rejection of any, or all, of the null contrast-specific hypotheses provides insight into the nature of the 'overall' treatment differences.

Despite the many attractions of orthogonal contrasts, less directed approaches are often advocated. Cox and Spjotvoll (1982) offered one such approach which aims to partition a set of means into groups through the use of standard F-tests. It is attractive because it avoids the use of any complicated probability calculations. Numerous alternative methods, often involving greater complexity, are frequently adopted. Scheffé (1953) developed a general approach to facilitate tests regarding any or all possible contrasts between treatment arms that ensures control of the experimental type I error rate arising from these multiple comparisons. A special case of Scheffé's approach involves performing all pairwise comparisons. Additional testing procedures designed to characterize the nature of the treatment differences include Duncan's multiple-range test (Duncan, 1955), the Newman–Keuls test (Newman, 1939; Keuls, 1952) and Tukey's (1949) multiple-comparison procedure, the last two being based on Studentized range statistics. Dunnett (1955) described a testing procedure designed to control the experimental type I error rate while allowing tests for each of $I - 1$ active treatments to the control; in his procedure these tests are not assumed to be predicated by rejection of some omnibus test for treatment differences.

Many other testing procedures have been developed specifically for multiarmed clinical trials, most sharing the objective of controlling the experimental type I error rate. See Bauer (1991) for a comprehensive list of references for a wide variety of alternative approaches. Further approaches are available when specific departures from the null hypothesis are of interest. For example, with active treatment arms consisting of increasing dosages of an experimental treatment, tests against ordered alternatives may be appropriate (Mau, 1988; Liu *et al.*, 1993). More general patterned alternatives can also be tested by using methods described by Hettmansperge and Norton (1987).

4.2. *Testing Selected Contrasts*

There are situations when there is understandable vagueness about what comparison might be of most interest. For example, in an epidemiological study to

investigate the relationship between HLA alleles and a specific disease, all pairwise comparisons may indeed be of interest (Prentice *et al.*, 1984). It is more difficult to imagine such non-specific hypotheses in a phase III clinical trial, however. Treatment arms are usually introduced into a trial for one or more particular reasons. For example, consider a trial in individuals infected with the human immunodeficiency virus which investigates the addition of a second antiretroviral agent (B) to an accepted standard (A). If, in addition, the effect of a proteinase inhibitor (C) were to be studied, four possible treatment arms would be A, $A+B$, $A+C$ and $A+B+C$. Suppose that the outcome of interest was progression to acquired immune deficiency syndrome (AIDS). Analyses of treatment effects in such a trial have been suggested based on generic multiple-comparison procedures involving all four treatment arms. However, it would seem that the particular questions of interest could be clearly specified as

(a) does the proteinase inhibitor add to the benefit of antiretroviral agents,
(b) does a second antiretroviral agent delay the onset of AIDS and
(c) does the use of agent B impact the effect of the proteinase inhibitor?

In the analysis of variance, an alternative to predicating further analyses on a test of an overall treatment effect is to specify the contrasts of interest at the design stage. This is implicitly done in the analysis of factorial designs when main effects are isolated but is also relevant to studies of single factors with greater than two levels. Prespecified contrasts ensure that comparisons are not chosen on the basis of possibly spurious observed differences in treatment means. Typically the contrasts are orthogonal and each is examined separately. In this case, there is usually minimal concern over the experimental type I error rate corresponding to these multiple tests.

Similarly, in a clinical trial the reasons for the introduction of the various treatment arms will usually define a small number of clinical questions of primary interest. These can be used to define comparisons, or contrasts, of interest. If the number of these contrasts is not excessive, then it is not reasonable to impose constraints solely to control the experimental type I error rate. Each contrast can be considered individually with separate p-values or type I error rates.

For example, suppose that H_{l0}: $\mathbf{c}_l'\theta = 0$ *versus* H_{lA}: $\mathbf{c}_l'\theta \neq 0$, $l = 1, \ldots, L$, represent L contrasts of interest with $\mathbf{c}_l = (c_{l1}, c_{l2}, \ldots, c_{lI})'$ a vector of constants and $\theta = (\theta_1, \theta_2, \ldots, \theta_I)'$ the parameter vector of interest. As with most tests of this form, a type I error corresponds to rejection of H_{l0} when in fact it is true. An experimental type I error rate could be defined as the frequency with which one or more null hypotheses would be rejected in the hypothetical infinite population of such trials, when they are all in fact true. It is not obvious that the aim of controlling the overall false positive error rate in the trial should have a major effect on the conclusions relating to questions which have specific interest in their own right. For clinical trials in which the specific contrasts of main interest are not identified *a priori* and all pairwise comparisons are planned, multiplicity adjustments and other approaches to controlling the type I error rate are a reasonable price to pay for not having thought out the primary objectives and performing an excessive number of tests.

To reiterate, the central difference between these two approaches involves the extent to which questions can be separately defined and the effect that this has on the interpretation of 'type I' errors. Given the unfocused nature of global tests of the null

hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_I$, it is very reasonable to proceed directly to hypothesis tests regarding contrasts in phase III clinical trials.

In the definition of prespecified contrasts, restriction to orthogonal contrasts ensures that the resulting test statistics are independent. This independence is often achieved, however, at the expense of the adoption of hypotheses of lesser clinical relevance. Thus this restriction is unattractive for practical reasons. A pragmatic approach is to limit the number of contrasts to one fewer than the number of treatment arms, the degrees of freedom associated with treatment. This is equivalent to restricting attention to well-defined regression models with non-singular design matrices. If the contrasts are not orthogonal then the interpretation of each is dependent on the others in the model, as in partial F-tests for the normal regression model. If the number of comparisons exceeds the number of degrees of freedom, then they are implicitly related to a singular design matrix and it is no longer clear that each comparison can be looked at separately. In this case, the view that the marginal tests can be used for inferring the treatment information from the trial is questionable.

It can be stressed that concern regarding any false positive conclusions may be of considerable importance if they are directly linked to decisions. For example, some phase II clinical trials are set up to select one or a few treatments for further investigation. In such a case, an experimental type I error rate is appropriately defined and a formal multiple-comparison procedure should be devised. This is essentially a 'selection' problem. In a phase III trial with multiple arms, however, the motivation for the different arms may not be selection but rather an examination of different aspects of treatment strategies. It is in these circumstances that a moderate number of clinically relevant hypotheses can be specified and that multiple-testing procedures need not be adopted.

5. REPEATED SIGNIFICANCE TESTS

Compared with that arising in a multiple-outcome or multiple-treatment clinical trial, a qualitatively different type of multiple-testing situation arises in the monitoring of clinical trials. Consider a two treatment arm trial and a single hypothesis of the form H_0: $\delta = 0$ which may be tested against a two-sided alternative H_A: $\delta \neq 0$ with $\delta = \delta_A$ the effect size of clinical interest. For ethical, practical and economic reasons it is often necessary to monitor data from moderate and large clinical trials to ensure that, as soon as there is sufficient evidence to claim that there is a beneficial treatment effect, patients are no longer randomized to treatments, and all eligible patients receive the preferred treatment. This can be achieved by performing repeated tests of significance based on accumulating data.

Armitage *et al.* (1969) pointed out that such repeated tests on accumulating data also lead to an inflation of the type I error. In response to this Armitage *et al.* (1969) developed sequential designs directed at controlling the overall type I error rate while facilitating continuous testing of the accumulating data. This approach was subsequently generalized by Pocock (1977), O'Brien and Fleming (1979) and others (e.g. Jennison and Turnbull (1989)) to facilitate repeated tests of significance based on equal increments of information time (Lan and DeMets, 1989), i.e. after equal numbers of patients have been randomized. In these contexts, given that a single well-defined hypothesis is to be tested, the interpretation of the associated type I

error, and the associated type I error rate, is straightforward. Therefore it is very appropriate to control this error rate and to adopt formal group sequential procedures with this in mind.

Sequential monitoring is not limited to two-arm single-response clinical trials, however, and there is a growing number of references on the use of sequential procedures in more complex clinical trial designs including those with multiple outcomes (Tang *et al.*, 1989; Lin, 1991; Farewell and Cook, 1992; Cook and Farewell, 1995; Cook, 1994, 1996) and multiple treatment arms (Lin and Liu, 1992; Hughes, 1993; Proschan *et al.*, 1994). Although some technical complexity is introduced with these more complicated designs, the general issues raised in Sections 3 and 4 are still relevant. For example, it may be appropriate, in some trials with a select number of well-defined hypotheses, to focus on monitoring strategies which control marginal error rates at standard levels. Cook (1996) describes a sequential procedure for monitoring more than one response to treatment with this in mind. Similarly, the nature of the hypotheses of primary interest should play a role in designing a monitoring procedure for trials with multiple treatment arms.

6. DISCUSSION

This paper has raised some issues which we feel warrant further discussion in the context of clinical trials. Primarily, the issues are non-technical and indeed there may be some concern that they may be overlooked if attention is focused only on technical aspects of multiple-testing procedures. In particular, a concern is that testing strategies are frequently adopted with the aim of controlling the experimental type I error rate without considering how this relates to the questions of main interest.

A motivation for much of the discussion has been the view that a clinical trial is not primarily a decision-making process, but rather a scientific experiment with, of course, an ethical design. Although an experiment will influence subsequent behaviour, the dependence of this behaviour on the evidential results of the trial may not be easily prespecified. The strength of the evidence regarding various scientific questions may have a major effect. Thus, the utilization of marginal test results and marginal p-values as inputs for a process of inductive inference is more consistent with this approach. Furthermore, the process of inductive behaviour implied by the Neyman–Pearson framework is somewhat unrealistic given the wide variety of other factors that will influence clinical decision-making regarding an experimental treatment. The simple fact that treatment recommendations are often based on both clinical and statistical significance indicates that statistical evidence is not sufficient in itself to influence behaviour. It is worth reiterating, however, that the Neyman–Pearson approach provides a useful framework for designing these clinical experiments.

A closely related issue, which has not been discussed in any detail, is the role of estimation in the interpretation of trial results. The form of the discrepancy measure given in Section 3.1 was chosen to illustrate the link between test statistics and pivotal quantities. The multiplicity issues raised here can be re-expressed in a possibly more compelling manner in terms of the interval estimates arising from such pivotal quantities. Consider the multiple-outcome scenario of Section 3 and suppose that five response variables are identified as of interest. Although the notion of an experimental type I error rate may be alluring in the hypothesis testing context, control of this error rate corresponds directly to control of the simultaneous coverage probability of five joint confidence intervals. The difficulty in interpreting such a set of joint confidence intervals (in practical terms, we can interpret three such interval estimates jointly, at best) reflects the difficulty in ascribing meaning to the experimental type I error rate.

We have deliberately not discussed issues surrounding subgroup analyses and exhaustive secondary analyses undertaken after the completion of the study. The exploratory nature of these analyses introduces aspects beyond the discussion of this paper.

An obvious consequence of controlling marginal error rates at nominal levels is an inflation of the 'experimental' type I error rate. This consequence is of little importance if there is little interest in the type of errors which motivates it. One result of this view is a potential relaxation of the frequent sharp distinction been primary and important secondary outcomes. In particular, with a small number of outcomes, designs based on several outcomes can be adopted with appropriate sample size calculations. However, when a very large number of responses are necessary to be recorded to characterize adequately some disease state that is difficult to define, then global test statistics or summary measures such as those described by O'Brien (1984) should be adopted. Note though that the planning of a study based on this analysis is difficult as it requires a specification of a complete covariance matrix.

The suggestion that decisions regarding the use of multiplicity adjustments should be made based on whether or not the hypotheses are well defined, of clinical importance and are prespecified means that an element of subjectivity remains. One guideline in the context of multiarmed trials is that, as the number of treatment comparisons increases, the analysis takes on the flavour of a selection problem, suggesting that traditional multiple-comparison procedures may be appropriate. In phase III trials, however, usually a relatively few treatment arms are introduced for particular reasons. In this case, specific tests should be formulated with these reasons in mind. This highlights the importance of good communication between clinical investigators and statisticians at all stages of the investigation.

One contribution to the increase of trials with multiple outcomes has been a recent trend towards the more formal use of quality-of-life measures and health economic measures in the determination of the value of treatments. These measures tend not to be restricted to a few outcomes and so a trial may result with a large number of responses identified at the design stage. One possible strategy in this case is to recognize that they are effectively grouped according to various characteristics (e.g. quality-of-life dimensions), and to construct summary measures or global statistics within these groups. The discussion of this paper would be most relevant to the joint consideration of these summary measures or global statistics. A related application of marginal procedures arises from recognizing the difficulty in combining quality-of-life measures with more difficult clinical outcomes such as survival. Previous attempts at specifying relevant summary measures such as quality-adjusted life years (Torrance, 1987) have met with some criticism (Cox *et al.*, 1992), whereas the use of global test statistics in this context seems generally inappropriate.

The multiplicity issues discussed here have been considered in the context of phase III clinical trials. They are not new, however, and have parallels in the general design literature. For multiple outcomes and multiarmed trials, the concept of type I error

rates has been re-examined. The control of marginal rather than experimental error rates requires that the precise nature of a small number of well-defined hypotheses be specified in advance to ensure that 'focusing' on marginal test results is not driven by the test results themselves. However, with this restriction, there are many scenarios in which multiplicity adjustments need not be adopted to control experimental type I error rates. The central idea behind this assertion is that, for well-defined null and alternative hypotheses, we have the capacity to interpret test results marginally and to draw inferences accordingly. The concern is that testing strategies are frequently adopted to control the overall error rate at the expense of obscuring and losing focus of the clinical questions of main interest. To reiterate Cox's (1965) comment, the simultaneous correctness of many statements does not necessarily need to be considered when focusing on a particular response.

ACKNOWLEDGEMENTS

This work was supported by the Societal Institute for Mathematical Sciences through National Institute on Drug Abuse grant DA04722 and by the Natural Sciences and Engineering Research Council of Canada. We would like to thank Dr Charlie Dunnett for comments on an earlier draft of this paper.

REFERENCES

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969) Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A, **132**, 235–244.
Bauer, P. (1991) Multiple testing in clinical trials. *Statist. Med.*, **10**, 871–890.
Cook, R. J. (1994) Interim monitoring of bivariate response using repeated confidence intervals. *Contr. Clin. Trials*, **15**, 187–200.
——(1996) Coupled error spending functions for parallel bivariate sequential tests. *Biometrics*, to be published.
Cook, R. J. and Farewell, V. T. (1995) Guidelines for monitoring efficacy and toxicity responses in clinical trials. *Biometrics*, **50**, 1146–1152.
Cox, D. R. (1965) A remark on multiple comparison methods. *Technometrics*, **7**, 223–224.
——(1977) The role of significance tests. *Scand. J. Statist.*, **4**, 49–70.
Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Spiegelhalter, D. J. and Jones, D. R. (1992) Quality-of-life assessment: can we keep it simple (with discussion)? *J. R. Statist. Soc.* A, **155**, 353–393.
Cox, D. R. and Spjotvoll, E. (1982) On partitioning means into groups. *Scand. J. Statist.*, **9**, 147–152.
Diddle, A. W., O'Connor, K. A. and Pearce, R. L. (1953) Evaluation of bed rest in threatened abortion. *Obstetr. Gyn.*, **2**, 63–67.
Duncan, O. B. (1955) Multiple range and multiple *F*-tests. *Biometrics*, **11**, 1–42.
Dunnett, C. (1955) A multiple comparisons procedure for comparing several treatments with a control. *J. Am. Statist. Ass.*, **50**, 1096–1121.
Farewell, V. T. and Cook, R. J. (1992) Comment: Evaluating therapeutic interventions: some issues and experiences. *Statist. Sci.*, **7**, 446–448.
Fisher, R. A. (1946) *Statistical Methods for Research Workers*. New York: Hafner.
——(1955) Statistical methods and scientific induction. *J. R. Statist. Soc.* B, **17**, 69–78.
——(1960) Scientific thought and the refinement of human reason. *J. Oper. Res. Soc. Jpn*, **3**, 1–10.
——(1971) *The Design of Experiments*. New York: Hafner.
——(1973) *Statistical Methods for Scientific Inference*. New York: Hafner.
Fitzpatrick, R., Bury, M., Frank, A. and Donnelly, T. (1987) Problems in the assessment of outcome in a back pain clinic. *Int. Disblty Stud.*, **9**, 161–165.
Fleming, T. R. (1990) Evaluation of active control trials in AIDS. *J. Acq. Immune Def. Synd.*, **3**, S82–S87.

Friend, P., Hale, G., Waldman, S., Gore, S., Thiru, S., Joysey, V., Evans, D. B. and Calne, R. Y. (1994) Campath-1 prophylactic use after kidney transplantation—a randomized controlled trial. To be published.
Gupta, S. S. (1963) Probability integrals of multivariate normal and multivariate *t*. *Ann. Math. Statist.*, **34**, 792–828.
Hettmansperge, T. P. and Norton, R. M. (1987) Tests for patterned alternatives in *K*-sample problems. *J. Am. Statist. Ass.*, **82**, 292–299.
Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.
Hoge, C. W., Shlim, D. R., Ghimire, M., Rabold, J. G., Pandey, P., Walch, A., Rajah, R., Gaudio, P. and Echeverria, P. (1995) *Lancet*, **345**, 691–693.
Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65–70.
Hoppe, F. M. (1993) *Multiple Comparisons, Selection, and Applications in Biometry*. New York: Dekker.
Hughes, M. H. (1993) Stopping guidelines for clinical trials with multiple treatments. *Statist. Med.*, **12**, 901–915.
Jennison, C. and Turnbull, B. W. (1989) Interim analyses: the repeated confidence interval approach (with discussion). *J. R. Statist. Soc.* B, **51**, 305–361.
Keuls, M. (1952) The use of the studentized range in connection with an analysis of variance. *Euphytica*, **1**, 112–122.
Lan, K. K. G. and DeMets, D. L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
——(1989) Group sequential procedures: calendar versus information time. *Statist. Med.*, **8**, 1191–1198.
Lehman, E. (1993) The Fisher, Neyman–Pearson theories of testing hypotheses: one theory or two? *J. Am. Statist. Ass.*, **88**, 1242–1249.
Lin, D. Y. (1991) Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika*, **78**, 123–131.
Lin, D. Y. and Liu, P. Y. (1992) Nonparametric sequential tests against ordered alternatives in multi-armed clinical trials. *Biometrika*, **79**, 420–427.
Liu, P. Y., Green, S., Wolf, M. and Crowley, J. (1993) Testing against ordered alternatives for censored survival data. *J. Am. Statist. Ass.*, **88**, 153–160.
Mau, J. (1988) A generalization of a nonparametric test for stochastically ordered distributions to censored survival data. *J. R. Statist. Soc.* B, **50**, 403–412.
Newman, D. (1939) The distribution of range in samples from a normal population, expressed as an independent estimate of standard deviation. *Biometrika*, **31**, 20–30.
Neyman, J. and Pearson, E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* A, **20**, 170–240, 263–294.
——(1933) On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. Lond.* A, **231**, 289–337.
North American Symptomatic Carotid Endarterectomy Trial Collaborators (1991) Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. *New Engl. J. Med.*, **325**, 445–453.
O'Brien, P. C. (1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.
O'Brien, P. C. and Fleming, T. R. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
Pocock, S. J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.
Pocock, S. J., Geller, N. L. and Tsiatis, A. A. (1987) The analysis of multiple endpoints in clinical trials. *Biometrics*, **43**, 487–498.
Prentice, R. L., Storb, R., Brown, K. S. and Mason, M. W. (1984) HLA and disease: relative risk regression methods and multiple testing considerations. *Biometrics*, **40**, 653–662.
Proschan, M. A., Follmann, D. A. and Geller, N. L. (1994) Monitoring multi-armed trials. *Statist. Med.*, **13**, 1441–1452.
Scheffé, H. (1953) A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.
Simes, R. J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Tang, D. I., Gnecco, C. and Geller, N. L. (1989) Design of group sequential clinical trials with multiple endpoints. *J. Am. Statist. Ass.*, **84**, 776–779.
Torrance, G. W. (1987) Utility approach to measuring health related quality of life. *J. Chron. Dis.*, **40**, 593–600.
Tukey, J. (1949) Comparing individual means in the analysis of variance. *Biometrics*, **5**, 99–114.
Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989) Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. *J. Am. Statist. Ass.*, **84**, 1065–1073.

J. R. Statist. Soc. A (1996)
159, *Part* 1, *pp.* 111–123

What's in a Mean? — an Examination of the Inconsistency between Men and Women in Reporting Sexual Partnerships

By J. WADSWORTH†,
St Mary's Hospital Medical School, London, UK

A. M. JOHNSON,
University College School of Medicine, London, UK

K. WELLINGS and J. FIELD
London School of Hygiene and Tropical Medicine, UK
Social and Community Planning Research, London, UK

[Received October 1994. Final revision July 1995]

SUMMARY

The mean rate of partner change is a key variable used in mathematical models of the transmission dynamics of sexually transmitted diseases and human immunodeficiency virus. This paper uses data from the British National Survey of Sexual Attitudes and Lifestyles to explore the consistency of responses given by men and women to questions about sexual life style. In common with other surveys of this nature, men report a higher mean number of heterosexual partners than do women. Possible explanations for such a finding are given in this paper and we show that the discrepancies can be reduced by making certain assumptions. The importance of the age mixing of sexual partnerships is highlighted.

Keywords: MEMORY ERROR; REPORTING BIAS; SEXUAL BEHAVIOUR

1. INTRODUCTION

The epidemic of the human immunodeficiency virus (HIV) has led to a renewed interest in epidemic modelling. One of the key elements of many of these models of HIV and other sexually transmitted agents is the mean rate of partner change (Hethcote and Yorke, 1984; Cox, 1989; Anderson *et al.*, 1986), which is usually estimated by the mean number of partners per unit time plus an element which reflects the variance of the mean, such as the variance:mean ratio. The National Survey of Sexual Attitudes and Lifestyles can provide such estimates, but it is important that the data are examined critically for reliability and evidence of bias.

1.1. *Accuracy*

The National Survey of Sexual Attitudes and Lifestyles differs little from other population surveys on sensitive topics though the problem of validating data on sexual behaviour may be more intractable. Self-reported data only are available and indirect measures must be used to assess their reliability (Wadsworth and Johnson, 1991). The accuracy of responses will depend on many factors, some of which can be modified through appropriate study design (Belson, 1981; Catania *et al.*, 1990). But the extent of social acceptability bias needs to be estimated and explored. Such bias

†*Address for correspondence*: Department of Epidemiology and Public Health, St Mary's Hospital Medical School, Praed Street, London, W2 1PG, UK.
E-mail: jhw3o@ic.ac.uk

0035–9238/96/159111