ELSEVIER

Journal of Clinical Epidemiology 54 (2001) 343–349

Journal of Clinical Epidemiology

# Adjusting for multiple testing—when and how?

Ralf Bender[a,]*, Stefan Lange[b]

[a]Institute of Epidemiology and Medical Statistics, School of Public Health, University of Bielefeld, P.O. Box 100131, D-33501 Bielefeld, Germany
[b]Department of Medical Informatics, Biometry and Epidemiology, Ruhr-University of Bochum, D-44780 Bochum, Germany

Received 9 September 1999; received in revised form 31 July 2000; accepted 2 August 2000

## Abstract

Multiplicity of data, hypotheses, and analyses is a common problem in biomedical and epidemiological research. Multiple testing theory provides a framework for defining and controlling appropriate error rates in order to protect against wrong conclusions. However, the corresponding multiple test procedures are underutilized in biomedical and epidemiological research. In this article, the existing multiple test procedures are summarized for the most important multiplicity situations. It is emphasized that adjustments for multiple testing are required in confirmatory studies whenever results from multiple tests have to be combined in one final conclusion and decision. In case of multiple significance tests a note on the error rate that will be controlled for is desirable.

*Keywords:* Multiple hypotheses testing; *P* value; Error rates; Bonferroni method; Adjustment for multiple testing; UKPDS

## 1. Introduction

Many trials in biomedical research generate a multiplicity of data, hypotheses, and analyses, leading to the performance of multiple statistical tests. At least in the setting of confirmatory clinical trials the need for multiple test adjustments is generally accepted [1,2] and incorporated in corresponding biostatistical guidelines [3]. However, there seems to be a lack of knowledge about statistical procedures for multiple testing. Recently, some authors tried to establish that the statistical approach of adjusting for multiple testing is unnecessary or even inadequate [4–7]. However, the main arguments against multiplicity adjustments are based upon fundamental errors in understanding of simultaneous statistical inference [8,9]. For instance, multiple test adjustments have been equated with the Bonferroni procedure [7], which is the simplest, but frequently also an inefficient method to adjust for multiple testing.

The purpose of this article is to describe the main concept of multiple testing, several kinds of significance levels, and the various situations in which multiple test problems in biomedical research may occur. A nontechnical overview is given to summarize in which cases and how adjustments for multiple hypotheses tests should be made.

## 2. Significance tests, multiplicity, and error rates

If one significance test at level $\alpha$ is performed, the probability of the type 1 error (i.e., rejecting the individual null hypothesis although it is in fact true) is the *comparisonwise error rate* (CER) $\alpha$, also called *individual level* or *individual error rate*. Hence, the probability of not rejecting the true null hypothesis is $1 - \alpha$. If $k$ independent tests are performed, the probability of not rejecting all $k$ null hypotheses when in fact all are true is $(1 - \alpha)^k$. Hence, the probability of rejecting at least one of the $k$ independent null hypotheses when in fact all are true is the *experimentwise error rate* (EER) under the complete null hypothesis $\text{EER} = 1 - (1 - \alpha)^k$, also called *global level*, or *familywise error rate* (considering the family of $k$ tests as one experiment). If the number $k$ of tests increases, the EER also increases. For $\alpha = 0.05$ and $k = 100$ tests EER amounts to 0.994. Hence, in testing 100 independent true null hypotheses one can almost be sure to get at least one false significant result. The expected number of false significant tests in this case is $100 \times 0.05 = 5$. Note that these calculations only hold if the $k$ tests are *independent*. If the $k$ tests are correlated no simple formula for the EER exists, because EER depends on the correlation structure of the tests.

Frequently, the global null hypothesis, that all individual null hypotheses are true simultaneously, is of limited interest to the researcher. Therefore, procedures for simultaneous statistical inference have been developed that control the *maximum experimentwise error rate* (MEER) under any complete or partial null hypothesis, also called *multiple*

* Corresponding author. Tel.: +49 521 106-3803; fax: +49 521 106-6465.

*E-mail address:* Ralf.Bender@uni-bielefeld.de (R. Bender)

0895-4356/01/$ – see front matter © 2001 Elsevier Science Inc. All rights reserved.
PII: S0895-4356(00)00314-0

*level*, or *familywise error rate in a strong sense*. The MEER is the probability of rejecting falsely at least one true individual null hypothesis, irrespective of which and how many of the other individual null hypotheses are true. A multiple test procedure that controls the MEER also controls the EER but not vice versa [10]. Thus, the control of the MEER is the best protection against wrong conclusions and leads to the strongest statistical inference.

The application of multiple test procedures enables one to conclude which tests are significant and which are not, but with control of the appropriate error rate. For example, when three hypotheses A, B, C are tested and the unadjusted $P$ values are $P_A = 0.01$, $P_B = 0.04$, and $P_C = 0.10$, the Bonferroni correction would lead to the adjusted $P$ values $P_A = 0.03$, $P_B = 0.12$, and $P_C = 0.30$. From this result we can conclude that test A is significant and tests B and C are not significant by controlling the MEER of 0.05.

## 3. When are adjustments for multiple tests necessary?

A simple answer to this question is: If the investigator only wants to control the CER, an adjustment for multiple tests is unnecessary; if the investigator wants to control the EER or MEER, an adjustment for multiple tests is strictly required. Unfortunately, there is no simple and unique answer to when it is appropriate to control which error rate. Different persons may have different but nevertheless reasonable opinions [11,12]. In addition to the problem of deciding which error rate should be under control, it has to be defined first which tests of a study belong to one experiment. For example, consider a study in which three different new treatments (T1, T2, T3) are compared with a standard treatment or control (C). All six possible pairwise comparisons (T1 vs. C, T1 vs. T2, T1 vs. T3, T2 vs. C, T2 vs T3, T3 vs. C) can be regarded as one experiment or family of comparisons. However, by defining the comparisons of the new treatments with the control (T1 vs. C, T2 vs. C, T3 vs. C) as the main goal of the trial and the comparisons of the new treatments among each other (T1 vs. T2, T1 vs. T3, T2 vs. T3) as secondary analysis, this study consists of two experiments of connected comparisons. In this case it may be appropriate to perform separate multiplicity adjustments in each experiment. In general, we think it is logical that the MEER should be under control when the results of a well-defined family of multiple tests should be summarized in one conclusion for the whole experiment. For example, if each new treatment is significantly different from the standard treatment, the conclusion that all three treatments differ from the standard treatment should be based upon an adequate control of the MEER. Otherwise the type 1 error of the final conclusion is not under control, which means that the aim of significance testing is not achieved.

Such a rigorous proceeding is strictly required in confirmatory studies. A study is considered as confirmatory if the goal of the trial is the definitive proof of a predefined key hypothesis for final decision making. For such studies a good predefined statistical analysis plan is required. A clear prespecification of the multiple hypotheses and their priorities is quite important. If it is possible to specify one clear primary hypothesis there is not multiplicity problem. If, however, the key hypothesis is proved by means of multiple significance tests, the use of multiple test procedures is mandatory.

On the other hand, in exploratory studies, in which data are collected with an objective but not with a prespecified key hypothesis, multiple test adjustments are not strictly required. Other investigators hold an opposite position that multiplicity corrections should be performed in exploratory studies [7]. We agree that the multiplicity problem in exploratory studies is huge. However, the use of multiple test procedures does not solve the problem of making valid statistical inference for hypotheses that were generated by the data. Exploratory studies frequently require a flexible approach for design and analysis. The choice and the number of tested hypotheses may be data dependent, which means that multiple significance tests can be used only for descriptive purposes but not for decision making, regardless of whether multiplicity corrections are performed or not. As the number of tests in such studies is frequently large and usually a clear structure in the multiple tests is missing, an appropriate multiple test adjustment is difficult or even impossible. Hence, we prefer that data of exploratory studies be analyzed without multiplicity adjustment. "Significant" results based upon exploratory analyses should clearly be labeled as exploratory results. To confirm these results the corresponding hypotheses have to be tested in further confirmatory studies.

Between the two extreme cases of strictly confirmatory and strictly exploratory studies there is a wide range of investigations representing a mixture of both types. The decision whether an analysis should be made with or without multiplicity adjustments is dependent on "the questions posed by the investigator and his purpose in undertaking the study" [13]. Whatever the decision is, it should clearly be stated why and how the chosen analyses are performed, and which error rate is controlled for.

In the following, we consider the case of a confirmatory study with a clear prespecified key question consisting of several hypotheses analyzed by multiple significance tests. These tests represent one experiment consisting of a family of connected significance tests. For a valid final conclusion an appropriate multiplicity adjustment should be made. We present a short nontechnical overview of statistical procedures for multiple test adjustment. More technical and comprehensive overviews can be found elsewhere [10,14–16].

## 4. General procedures for multiple test adjustments

### *4.1. General procedures based upon P values*

The simplest multiple test procedure is the well-known Bonferroni method [17]. Of $k$ significance tests, those accepted as statistically significant have $P$ values smaller than

$\alpha/k$, where $\alpha$ is the MEER. Adjusted $P$ values are calculated by $k \times P_i$, where $P_i$ for $i = 1, \dots, k$ are the individual unadjusted $P$ values. In the same manner Bonferroni adjusted confidence intervals can be constructed by dividing the multiple confidence level with the number of confidence intervals. The Bonferroni method is simple and applicable in essentially any multiple test situation. However, the price for this simplicity and universality is low power. In fact, the Bonferroni method is frequently not appropriate, especially if the number of tests is large. Bonferroni corrections should only be used in cases where the number of tests is quite small (say, less than 5) and the correlations among the test statistics are quite low.

Fortunately, there are a number of improvements of the Bonferroni method [2,16,18], such as the well-known Holm procedure [19,20]. Some of these modified Bonferroni methods represent stepwise procedures based upon the closed testing procedure introduced by Marcus et al. [21], which is a general principle leading to multiple tests controlling the multiple level [10]. A general algorithm for obtaining adjusted $P$ values for any closed test procedure is outlined by Wright [16]. While some of these methods are quite complex, the Holm method is just as simple and generally applicable as the Bonferroni method, but much more powerful [16,18].

### 4.2. Resampling-based procedures

Despite being more powerful than the simple Bonferroni method, the modified Bonferroni methods still tend to be conservative. They make use of the mathematical properties of the hypotheses structure, but they do not take the correlation structure of the test statistics into account. One approach that uses the information of dependencies and distributional characteristics of the test statistics to obtain adjusted $P$ values is given by resampling procedures [22]. For highly correlated tests, this approach is considerably more powerful than the procedures discussed above. However, the price for the gain of power is that the resampling-based procedures are computer intensive. PROC MULTTEST of SAS offers resampling-based adjusted $P$ values for some frequently used significance tests [22,23].

## 5. Special procedures for multiple test adjustments

One main advantage of the general multiple test procedures based upon $P$ values is that they are universally applicable to different types of data (continuous, categorical, censored) and different test statistics (e.g., $t$, $\chi^2$, Fisher, logrank). Naturally, these procedures are unspecific and special adjustment procedures have been developed for certain questions in specific multiplicity situations.

### 5.1. More than two groups

One area in which multiplicity adjustment has a long history is the comparison of the means of several groups in analysis of variance (ANOVA) [24]. For this application a number of procedures exist. The most well-known methods, which are frequently implemented in ANOVA procedures of statistical software packages, are the following. The simultaneous test procedures of *Scheffé* and *Tukey* can also be used to calculate simultaneous confidence intervals for all pairwise differences between means. The method of *Dunnett* can be used to compare several groups with a single control. In contrast to these single-step procedures, multiple stage tests are in general more powerful but give only homogenous sets of treatment means but no simultaneous confidence intervals. The most well-known multiple stage tests are the procedures of *Duncan*, *Student–Newman–Keuls* (SNK), and *Ryan–Einot–Gabriel–Welsch* (REGW). These procedures, with the exception of Duncan, preserve the MEER, at least in balanced designs. Which of these tests are appropriate depends on the investigator's needs and the study design. In short, if the MEER should be under control, with no confidence intervals needed and a balanced design, then the REGW procedure can be recommended. If confidence intervals are desirable or the design is unbalanced, then the Tukey procedure is appropriate. In case of ordered groups (e.g., dose finding studies), procedures for specific ordered alternatives can be used with a substantial gain in power [10]. More detailed overviews of multiple test procedures for the comparison of several groups are given elsewhere [16,25–27]. Multiple comparison procedures for some nonparametric tests are also available [28].

In the frequent case of three groups the principle of closed testing leads to the following simple procedure that keeps the multiple level $\alpha$ [10]. At first, test the global null hypothesis that all three groups are equal by a suitable level $\alpha$ test (e.g., and $F$ test or the Kruskal–Wallis test). If the global null hypothesis is rejected proceed with level $\alpha$ tests for the three pairwise comparisons (e.g., $t$ tests or Wilcoxon rank sum tests).

### 5.2. More than one endpoint

The case of multiple endpoints is one of the most common multiplicity problems in clinical trials [29,30]. There are several possible strategies to deal with multiple endpoints. The simplest approach, which should always be considered first, is to specify a *single primary endpoint*. This approach makes adjustments for multiple endpoints unnecessary. However, all other endpoints are then subsidiary and results concerning secondary endpoints can only have an exploratory rather than a confirmatory interpretation. The second possibility is to combine the outcomes in *one aggregated endpoint* (e.g., a summary score for quality of life data or the time to the first event in the case of survival data). The approach is adequate only if one is not interested in the results of the individual endpoints. Thirdly, for significance testing *multivariate methods* [e.g., multivariate analysis of variance (MANOVA) or Hotelling's $T^2$ test] and *global test statistics* developed by O'Brien [31] and extended by Pocock et al. [32] can be used. Exact tests suitable for a large number of endpoints and small sample size have been developed by Läuter [33]. All these methods provide an overall assessment of effects in terms of statistical significance but offer no estimate of the magnitude of the effects. Again, information about the effects concerning the individual endpoints is lacking. In addition, Hotelling's $T^2$ test lacks power since it tests for unstructured alternative hypotheses, when in fact one is really interested in evidence from several outcomes pointing in the same direction [34]. Hence, in the case of several equally important endpoints for which individual results are of interest, multiple test adjustments are required, either alone or in combination with previously mentioned approaches. Possible methods to adjust for multiple testing in the case of multiple endpoints are given by the general adjustment methods based upon *P* values [35] and the resampling methods [22] introduced above. It is also possible to allocate different type 1 error rates to several not equally important endpoints [36,37].

### *5.3. Repeated measurements*

Methods to adjust for multiple testing in studies collecting repeated measurements are rare. Despite much recent work on mixed models [38,39] with random subject effects to allow for correlation of data, there are only few multiple comparison procedures for special situations. It is difficult to develop a general adjustment method for multiple comparisons in the case of repeated measurements since these comparisons occur for between-subject factors (e.g., groups), within-subject factors (e.g., time), or both. The specific correlation structure has to be taken into account, involving many difficulties. If only comparisons for between-subject factors are of interest, one possibility is to consider the repeated measurements as multiple endpoints and use one of the methods mentioned in the previous section. However, if the repeated measurements are ordered, this information is lost by using such an approach.

If repeated measurements are collected serially over time, the use of *summary measures* (e.g., area under the curve) to describe the response curves should be considered [40,41]. The analysis takes the form of a two-stage method where, in the first step, suitable summary measures for each response curve are calculated , and in the second step, these summary measures are analyzed by using the approaches discussed above. The choice of an adequate approach in the second stage depends on the number of groups to be compared and the number of summary measures to be analyzed. Only in the case of two groups and one summary measure as single primary endpoint does no multiplicity problem arise. To compare response curves between groups, Zerbe and Murphy have developed an extension of the Scheffé method and a stepwise procedure to adjust for multiple testing [42]. There are also multiple comparison procedures for some nonparametric tests suitable for related samples [28].

### *5.4. Subgroup analyses*

The extent to which subgroup analyses should be undertaken and reported is highly controversial [43,44]. We will not discuss the full range of problems and issues related to subgroup analyses but focus on the multiplicity problem. If one is interested in demonstrating a difference in the magnitude of the effect size between subgroups, a statistical test of interaction is appropriate, although such tests generally have low power [45]. If it is the aim to show an effect in all (or in some) of a priori defined subgroups on the basis of existing hypotheses, an adjustment for multiple testing should be performed by using one of the general procedures based upon *P* values. If there are few nonoverlapping subgroups, a test within one subgroup is independent of a test within another subgroup. In this case, the use of the simple Bonferroni method is possible. Frequently, however, subgroup analyses are performed concerning subgroups that are defined a posteriori after data examination. In this case, the results have an exploratory character regardless of whether multiplicity adjustments are performed or not. For interpretation of such analyses one should keep in mind that the overall trial result is usually a better guide to the effect in subgroups than the estimated effect in the subgroups [46].

### *5.5. Interim analyses*

Interim analyses of accumulating data are used in long-term clinical trials with the objective to terminate the trial when one treatment is significantly superior to the other(s). Since repeated analyses of the data increase the type 1 error, multiplicity adjustments are required for the development of adequate stopping rules. A simple rule that may be sufficient in many trials is: if no more than 10 interim analyses are planned and there is one primary endpoint, then $P < .01$ can be used as criterion for stopping the trial, because the global level will not exceed .05 [47]. The disadvantage of this approach is that the final analysis has to be undertaken at a significance level considerably smaller than .05 (also .01). Another simple possibility is to be extremely cautious for stopping the trial early by using $P < .001$ for the interim analyses [48]. This approach covers any number of interim analyses and is so conservative that the final analysis can be conducted at the usual .05 level. A compromise between these approaches is to use the procedure developed by O'Brien and Fleming [49] with varying nominal significance levels for stopping the trial. Early interim analyses have more stringent significance levels while the final analysis is undertaken as close to the .05 level as possible. Overviews about recent developments in the field of interim monitoring of clinical trials are given elsewhere [50–56].

## 6. Discussion

The problem of multiple hypotheses testing in biomedical research is quite complex and involves several difficulties. Firstly, it is required to define which significance tests belong to one experiment; that means which tests should be used to make one final conclusion. Secondly, the particular error rate to be under control must be chosen. Thirdly, an appropriate method for multiple test adjustment has to be found that is applicable and feasible in the considered situation. Many multiple test procedures for standard situations have been developed, but in the practice of clinical and epidemiological trials, there are a lot of situations in which an adequate control of type 1 error is quite complex, especially if there are several levels of multiplicity (e.g., more than two groups *and* more than one endpoint *and* repeated measurements of each endpoint). Unfortunately, the level of complexity can be so high that it is impossible to make an adequate adjustment for multiple testing. For example, the UK Prospective Diabetes Study (UKPDS) [57] contains an enormous complexity regarding multiplicity. Considering only the four main UKPDS publications in the year 1998 (UKPDS 33,34,38,39) [58–61] (i.e., neglecting the interim analyses and multiple significance tests published in earlier and future articles) there are 2, 4 or 5 main treatment groups (dependent on the question), additional comparisons between specific medications (e.g., captopril vs. atenolol), approximately 50 endpoints (7 aggregated, 21 single, 8 surrogate, and 12 compliance endpoints), and subgroup analyses (e.g., regarding overweight patients).

Of course, for such a specific and complex design no adequate and powerful multiple test procedure exists. Although Bonferroni adjustments would be principally possible, they would allow only comparisons with *P* values below .00017 (.05/298) to be significant, as we counted 298 different *P* values in the four articles. Naturally, with nearly 300 tests the Bonferroni procedure has not enough power to detect any true effect and cannot be recommended here. The UKPDS group tried to account for multiplicity by calculating 99% confidence intervals for single endpoints [60]. This approach slightly reduces the risk of type 1 error, but for a confirmatory study this procedure is not an adequate solution since the main goal of a significance test, namely the control of the type 1 error to a given level, is not achieved. Moreover, although 99% confidence intervals were calculated, unadjusted *P* values were presented with the effect that they are interpreted with the usual 5% level of significance [62]. Hence, in the UKPDS no firm conclusions can be drawn from the significance tests as the actual global significance level exceeds 5% by a large and unknown amount.

To avoid such difficulties a careful planning of the study design is required, taking multiplicity into account. The easiest and best interpretable approach is to avoid multiplicity as far as possible. A good predefined statistical analysis plan and a prespecification of the hypotheses and their priorities will in general reduce the multiplicity problem. If multiplicity can not be avoided at all (e.g., because there are several equally important endpoints), the investigators should clearly define which hypotheses belong to one experiment and then adjust for multiple testing to achieve a valid conclusion with control of the type 1 error. In the UKPDS one could have defined the intensive versus the conventional treatment as the primary comparison, with the consequence that confirmatory statements concerning the different intensive treatments are impossible. Furthermore, one could have defined the aggregated endpoint “any diabetes-related endpoint” as the primary outcome, with the consequence that all other aggregated endpoints are subsidiary. By means of a closed testing procedure it would have been possible to perform tests concerning the single endpoints forming the primary aggregated outcome (e.g., blindness, death from hypoglycemia, myocardial infarction, etc.) by preserving the MEER. The number of confirmatory analyses would be drastically reduced, but the results would be interpretable.

A further problem we did not mention in detail concerns the type of research in which estimates of association can be obtained for a broad range of possible predictor variables. In such studies, authors may focus on the most significant of several analyses—a selection process that may bias the magnitude of observed associations (both point estimates and confidence intervals). One way to deal with this type of multiplicity problem is to demand reproduction of the observed associations and their magnitude in further independent trials. However, this ‘solution’ does not address the adjustment of significance levels. A data-driven analysis and presentation, also called ‘data dredging’ or ‘data fishing,’ can only produce exploratory results. It can be used to generate hypotheses but not to test and confirm them, regardless of whether multiplicity corrections are performed or not. Hence, the use of multiple test procedures cannot protect against the bias caused by data fishing.

In principal, there is an alternative approach to significance testing for analysis of data. Bayes methods differ from all the methods discussed above in minimizing the Bayes risk under additive loss rather than controlling type 1 error rates. From a Bayesian perspective control of type 1 error is not necessary to make valid inferences. Thus, the use of Bayes methods avoids some of the conceptual and practical difficulties involved with the control of type 1 error, especially in the case of multiplicity. Hence, Bayes methods are useful for some of the multiplicity situations discussed above. Examples are the monitoring of clinical trials [63] and the use of empirical Bayes methods for the analysis of a large number of related endpoints [64,65]. However, in this article we concentrate on classical statistical methods based upon significance tests. We started from the assumption that an investigator had decided to use significance tests for data analysis. For this case we tried to summarize the available corresponding procedures to adjust for multiple testing. Bayes methods—which do not provide adjustments of *P* values as they do not give *P* values at all—are not further discussed here.

In summary, methods to adjust for multiple testing are valuable tools to ensure valid statistical inference. They should be used in all confirmatory studies where on the basis of a clearly defined family of tests one final conclusion and decision will be drawn. In such cases the maximum experimentwise error rate under any complete or partial null hypothesis should be under control. While the simple Bonferroni method is frequently not appropriate due to low power, there are a number of more powerful approaches applicable in various multiplicity situations. These methods deserve wider knowledge and application in biomedical and epidemiological research.

## Acknowledgment

We thank Dr. Gernot Wassmer (Cologne, Germany) for his careful reading of the manuscript and his valuable comments.

## References

[1] Koch GG, Gansky SA. Statistical considerations for multiplicity in confirmatory protocols. Drug Inf J 1996;30:523–34.
[2] Sankoh AJ, Huque MF, Dubin N. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med 1997;16:2529–42.
[3] The CPMP Working Party on Efficacy of Medical Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medical products. Stat Med 1995;14:1659–82.
[4] Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990;1:43–6.
[5] Savitz DA, Olshan AF. Multiple comparisons and related issues in the interpretation of epidemiologic data. Am J Epidemiol 1995;142:904–8.
[6] Savitz DA, Olshan AF. Describing data requires no adjustment for multiple comparisons: a reply from Savitz and Olshan. Am J Epidemiol 1998;147:813–4.
[7] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236–8.
[8] Aickin M. Other method for adjustment of multiple testing exists [Letter]. BMJ 1999;318:127.
[9] Bender R, Lange S. Multiple test procedures other than Bonferroni's deserve wider use [Letter]. BMJ 1999;318:600–1.
[10] Bauer P. Multiple testing in clinical trials. Stat Med 1991;10:871–90.
[11] Thompson JR. Invited commentary: Re: "Multiple comparisons and related issues in the interpretation of epidemiologic data." Am J Epidemiol 1990;147:801–6.
[12] Goodman SN. Multiple comparisons, explained. Am J Epidemiol 1998;147:807–12.
[13] O'Brien PC. The appropriateness of analysis of variance and multiple comparison procedures. Biometrics 1983;39:787–94.
[14] Miller RG. Simultaneous statistical inference. New York: McGraw-Hill, 1966.
[15] Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: Wiley, 1987.
[16] Wright SP. Adjusted p-values for simultaneous inference. Biometrics 1992;48:1005–13.
[17] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310:170.
[18] Levin B. Annotation: on the Holm, Simes, and Hochberg multiple test procedures. Am J Public Health 1996;86:628–9.
[19] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6:65–70.
[20] Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am J Public Health 1996;86:726–8.
[21] Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976;63:655–60.
[22] Westfall PH, Young SS. Resampling-based multiple testing. New York: Wiley, 1993.
[23] Westfall PH, Young SS. Reader reaction: on adjusting P-values for multiplicity. Biometrics 1993;49:941–5.
[24] Altman DG, Bland JM. Comparing several groups using analysis of variance. BMJ 1996;312:1472–3.
[25] Godfrey K. Comparing means of several groups. N Engl J Med 1985;313:1450–6.
[26] Jaccard J, Becker MA, Wood G. Pairwise multiple comparison procedures: a review. Psychol Bull 1984;96:589–96.
[27] Seaman MA, Levin JR, Serlin RC. New developments in pairwise multiple comparisons: some powerful and practicable procedures. Psychol Bull 1991;110:577–86.
[28] Conover WJ. Practical nonparametric statistics. New York: Wiley, 1980.
[29] Pocock SJ. Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation. Contr Clin Trials 1997;18:530–45.
[30] Zhang J, Quan H, Ng J. Some statistical methods for multiple endpoints in clinical trials. Contr Clin Trials 1997;18:204–21.
[31] O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics 1984;40:1079–87.
[32] Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. Biometrics 1987;43:487–98.
[33] Läuter J. Exact t and F tests for analyzing studies with multiple endpoints. Biometrics 1996;52:964–70.
[34] Follmann D. Multivariate tests for multiple endpoints in clinical trials. Stat Med 1995;14:1163–75.
[35] Lehmacher W, Wassmer G, Reitmeir P. Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. Biometrics 1991;47:511–21.
[36] Moyé LA. P-value interpretation and alpha allocation in clinical trials. Ann Epidemiol 1998;8:351–7.
[37] Moyé LA. Alpha calculus in clinical trials: Considerations and commentary for the new millennium. Stat Med 2000;19:767–79.
[38] Cnaan A, Laird NM, Slasor P. Tutorial in biostatistics: using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. Stat Med 1997;16:2349–80.
[39] Burton P, Gurrin L, Sly P. Tutorial in biostatistics: extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med 1998;17:1261–91.
[40] Matthews JNS, Altman DG, Campbell MJ, Royston P. Analysis of serial measurements in medical research. BMJ 1990;300:230–5.
[41] Senn S, Stevens L, Chaturvedi N. Tutorial in biostatistics: repeated measures in clinical trials: simple strategies for analysis using summary measures. Stat Med 2000;19:861–77.
[42] Zerbe GO, Murphy JR. On multiple comparisons in the randomization analysis of growth and response curves. Biometrics 1986;42:795–804.
[43] Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. Ann Intern Med 1992;116:78–84.
[44] Feinstein AR. The problem of cogent subgroups: a clinicostatistical tragedy. J Clin Epidemiol 1998;51:297–9.
[45] Buyse ME. Analysis of clinical trial outcomes: some comments on subgroup analyses. Contr Clin Trials. 1989;10:187S–94S.
[46] Yusuf S, Wittes J, Probstfield J, Tyroler A. Analysis and interpretation of treatment effects in subgroups of patients in randomised clinical trials. JAMA 1991;266:93–8.
[47] Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
[48] Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br J Cancer 1976;34:585–612.

[49] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–56.

[50] DeMets DL, Lan KKG. Overview of sequential methods and their applications in clinical trials. Commun Stat A 1984;13:2315–38.

[51] Geller NL, Pocock SJ. Interim analyses in randomized clinical trials: ramifications and guidelines for practioners. Biometrics 1987;43: 213–23.

[52] Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. Stat Sci 1990;5:299–317.

[53] Pocock SJ. Statistical and ethical issues in monitoring clinical trials. Stat Med 1993;12:1459–69.

[54] Lee JW. Group sequential testing in clinical trials with multivariate observations: a review. Stat Med 1994;13:101–11.

[55] Facey KM, Lewis JA. The management of interim analyses in drug development. Stat Med 1998;17:1801–9.

[56] Skovlund E. Repeated significance tests on accumulating survival data. J Clin Epidemiol 1999;52:1083–8.

[57] The UK Prospective Diabetes Study (UKPDS) Group. U.K. Prospective Diabetes Study (UKPDS): VIII. Study design, progress and performance. Diabetologia 1991;34:877–90.

[58] The UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). Lancet 1998;352:837–53.

[59] The UK Prospective Diabetes Study (UKPDS) Group. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). Lancet 1998; 352:854–65.

[60] The UK Prospective Diabetes Study (UKPDS) Group. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. BMJ 1998;317:703–13.

[61] The UK Prospective Diabetes Study (UKPDS) Group. Efficacy of atenolol and captopril in reducing risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 39. BMJ 1998;317: 713–20.

[62] de Fine Olivarius N, Andreasen AH. The UK Prospective Diabetes Study [Letter]. Lancet 1998;352:1933.

[63] Fayers PM, Ashby D, Parmar MKB. Tutorial in biostatistics: Bayesian data monitoring in clinical trials. Stat Med 1997;16:1413–30.

[64] Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987;43:671–81.

[65] Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. Epidemiology 1991;2:244–51.