The Likelihood as Statistical Evidence in Multiple Comparisons in Clinical Trials: No Free Lunch

Edward L. Korn* and **Boris Freidlin**

Biometric Research Branch, EPN-8129, National Cancer Institute, Bethesda, MD 20892-7434, USA

Received 2 December 2005, accepted 23 December 2005

Summary

The likelihood ratio summarizes the strength of statistical evidence for one simple pre-determined hypothesis versus another. However, it does not directly address the multiple comparisons problem. In this paper we discuss some concerns related to the application of likelihood ratio methods to several multiple comparisons issues in clinical trials, in particular, subgroup analysis, multiple variables, interim monitoring, and data driven choice of hypotheses.

Key words: Bayesian methods; Interim monitoring; Frequentist methods; Subgroup analysis; Subset analysis; Likelihood ratio.

1 Introduction

The use of likelihood functions has long been discussed and debated as a measure of statistical evidence (Hacking, 1965; Edwards, 1972). In particular, for assessing the evidence for a simple hypothesis A versus a simple hypothesis B based on data x, one calculates the likelihood ratio $f_A(x)/f_B(x)$, where $f_H(x)$ is the probability (or probability density function) of observing x under the hypothesis H. Values of the likelihood ratio greater than 1 suggest evidence for hypothesis A over B, with values ≥ 8 being suggested as fairly strong evidence and ≥ 32 strong evidence (Royall, 2000); similar cut-offs have been suggested by others (Jeffreys, 1961; Edwards, 1972; Kass and Raftery, 1995). A possible *raison d'être* for defining statistical evidence in this manner is as follows: if one had prior probabilities that each of the hypotheses were true, then from Bayes theorem the ratio of posterior probabilities of the hypotheses given the data is equal to the ratio of prior probabilities times the likelihood ratio (Edwards, 1972, p. 46). Alternatively, one might axiomatically derive the likelihood ratio as a measure of statistical evidence from desired criteria one would want such a measure to have (Good, 1985). Statistical evidence should not be confused with the propensity to believe a hypothesis or with a prescription to act in a certain way (Royall, 1997, pp. 3–5).

The likelihood plays a central role in statistical theory. However, some proponents of likelihood methods have been making their case in a way that may be interpreted as suggesting that use of the likelihood methods eliminates problems with multiple comparisons (Royall, 2000; Royall, 1997, pp. 111–116; Blume, 2002). In particular, it has been suggested in a recent tutorial (Blume, 2002, p. 2586) that in the subgroup analyses of clinical trials, "Likelihood based methods permit the construction of any number of subgroups and allow for the evaluation of their evidence without adjustment", and for interim monitoring of clinical trials (Blume, 2002, p. 2563), "Under this paradigm, re-examination of accumulating evidence is encouraged because (i) the likelihood ratio, unlike a p-value, is unaffected by the number of examinations and (ii) the probability of observing strongly misleading evidence is naturally low, even for study designs that re-examine the data with each new observation."

* Corresponding author: e-mail: korne@ctep.nci.nih.gov Phone: +1 301 402 0635, Fax: +1 301 402 0560

As multiple comparisons problems can be serious in clinical trials (as we will discuss below), it is useful to clarify some of these issues. We consider four areas in turn: subgroup analyses, multiple variables, interim monitoring, and data driven hypotheses. We conclude with a discussion of the Bayesian approach to these areas.

2 Subgroup Analyses

In a randomized clinical trial, eligible patients who are willing to participate are randomized between competing therapies. Subgroup analysis refers to analyzing the treatment effect on a subset of the randomized patients. The multiple comparisons problem arises when the treatment effect on many subgroups is examined – the effect will appear large and possibly statistically significant on the subgroup showing the largest treatment effect even if there is zero true treatment effect for all of the subgroups. Collins et al. (1987) give a humorous example in which treatment effect for patients with the astrological birth sign Scorpio appears to be larger than for other patients. A serious example is provided by the presentation of a recent AIDS vaccine trial in which overall the proportions infected were virtually the same ($191/3330 = 5.7\%$ for the vaccine versus $98/1679 = 5.8\%$ for the placebo) but restricting to combined race group "Black/Asian/other" the treatment effect was large ($12/327 = 3.7\%$ for the vaccine versus $17/171 = 9.9\%$ for the placebo) (Cohen, 2003a). Apparently, nine subgroup analyses were done on race, and the presentation did not make clear that there was no adjustment for the multiple comparisons when the p-values were given (Cohen, 2003b). Since clinical trials are used to develop treatment guidelines, it appears clear that some consideration for the multiple comparisons involved in subgroup analyses is necessary to keep ineffective treatments from becoming part of clinical practice.

We now consider a simple hypothetical example to demonstrate how likelihood-based methods accommodate subgroup analyses. Consider a clinical trial in which individuals with high blood pressure are randomized to either a new blood pressure lowering drug ("drug X") or placebo. An average drop in (systolic) blood pressure of 20 mg Hg would be considered clinically meaningful. Unfortunately, the results of the trial are negative, with the average drops in blood pressure for the drug and placebo groups being about the same. The principal investigator of the trial decides to do some data mining and discovers that in the subset of vegetarian post-menopausal women who were taking daily vitamins the effect of the treatment was dramatic and the largest seen in any of the numerous subsets he examined: The mean drop in blood pressure for individuals in this subset who were taking drug X was 33.8 with standard error of 10 and was 0.0 with standard error of 10 for those in the subset taking placebo. Using a normal-distribution model for the difference in means, the likelihood ratio of the hypothesis of a true difference of 20 versus the hypothesis of a true difference of 0 is 8. Incidentally, the 95% confidence interval (unadjusted for multiple comparisons) for the treatment effect is (6.1, 61.5).

Now consider another hypothetical trial of a cholesterol-lowering drug ("drug Y") versus placebo for individuals with high cholesterol, with an average drop of 20 mg/dL considered clinically meaningful. The principal investigator, consulting with the study statistician and study sponsors, decides to specify in the protocol before the trial begins a single subset for which he would like to do a separate analysis because he thinks the drug may be especially effective in that subset. This subset is vegetarian post-menopausal women who are taking daily vitamins. The results of the trial turn out to be identical to those of the previous trial; no effect of drug Y overall, but in the subset the average drops in cholesterol are 33.8 and 0.0 for the drug and placebo groups, respectively, both with standard error of 10. The likelihood ratio and confidence interval are obviously the same as in the previous trial, too.

It would be easy to infer from the quotes in the Introduction that these two trials could be reported in exactly the same way. That is, the likelihood ratios from the subgroup analysis could be presented with no mention of the fact that many subgroups were examined and only the one with the highest likelihood ratio being reported in the first case and a single pre-specified subgroup was examined in

the second case. If one defines the statistical evidence as being measured by the likelihood ratio, then the statistical evidence is the same concerning the pairs of hypotheses. (Since the means and standard errors for the subgroup in question are sufficient statistics for the parameter under consideration, the other data from the trial are irrelevant.) One could make the same argument using frequentist logic and report the confidence intervals for the subsets without adjustment for multiple comparisons; see Miller (1981, pp. 5–12) for a discussion of control of familywise error versus control of comparisonwise error. In terms of analysis, therefore, likelihood ratios handle multiple comparisons problems with subgroups about as well as unadjusted confidence intervals – they do not address the issue. We believe that something more is needed for the proper interpretation of subgroup analyses. One possibility is to use likelihood ratios in the context of a Bayesian analysis that accounts for multiple subgroups; see the Discussion.

Interestingly, in the *design* of a trial that is going to report likelihood ratios as the strength of statistical evidence, one could make adjustments for planned subgroup analyses. In the case of a single comparison, Royall (2000) has suggested choosing a sample size such that if the null hypothesis is true, the probability of observing misleading evidence (e.g., a likelihood ratio ≥ 8) is small. With multiple hypotheses, one could consider choosing a (larger) sample size so that the probability of observing misleading evidence for one or more of the hypotheses would be small. If this proposal was implemented, the result would be increases in sample sizes (and decreases in the probabilities of observing misleading evidence for individual comparisons) similar to those required by frequentist control of the familywise error.

3 Multiple Variables

In the context of a cancer clinical trial, consider a study to determine whether the gene expression of tumor cells is associated with the survival of the patients. The investigators perform a microarray analysis on pre-treatment biopsy specimens, and thereby acquires gene expression values for 10 000 genes for each patient. The investigators report the likelihood ratio for the single gene that is most associated with survival. They do not show the results for the other genes, or even mention that there were other genes assessed. They do this because they think no adjustment for multiple comparisons is required for likelihood methods. We believe that something more is need for the proper interpretation of this analysis involving 10 000 comparisons. Frequentist and Bayesian methods will pay a price for the multiple comparisons by requiring a larger effect to be seen for this single gene for the result to be considered interesting. We believe this is appropriate.

4 Interim Monitoring

Consider a randomized placebo controlled clinical trial designed to evaluate a new cholesterol lowering drug in a single analysis after a fixed number of patients have been observed. Assume that (1) cholesterol levels are normally distributed with known variance τ^2, and (2) the minimum clinically meaningful reduction in cholesterol is δ. Let hypothesis A be $\Delta = 0$ and hypothesis B be $\Delta = \delta$, where Δ is the amount by which the drug lowers cholesterol level relative to placebo (the difference between the 6-month post randomization and baseline cholesterol levels). For a classical frequentist design, one chooses the sample size and rejection region based on δ^2/τ^2 so that the probability that hypothesis A is rejected when it is true is low (the type I error) and the probability that it is accepted when hypothesis B is true is low (the type II error). It has been universally accepted that for ethical reasons ongoing trials should be monitored for strong indications of efficacy or harm. If interim looks at the data are used to possibly stop the trial early and reject hypothesis A, then the type I error will be increased over the fixed sample size design – this is the frequentist cost of the interim monitoring.

One can maintain the type I and II errors and retain interim monitoring by adjusting the stopping boundary and increasing the sample size. If this is not done, then one can stop trials too early and

draw incorrect conclusions. An example is given by the data monitoring of a trial of two drugs for AIDS in which an interim monitoring boundary was not crossed but if one had used the nominal significance to assess stopping the trial, an incorrect decision would have been made (Fleming et al., 1995). Another example is given by a trial of 5 years versus more than 5 years of tamoxifen therapy for breast cancer patients which was stopped early with the conclusion that no additional benefit was provided by more than five years of therapy. The conclusion was questioned in part because "public availability of the results has been influenced by the patterns that they suggest" (Peto, 1996). However, the effects of interim monitoring were considered in the stopping decision (Dignam et al., 1998).

Royall (1997, p. 53) proposed a likelihood-ratio-based design with fixed sample size chosen so that the probability of obtaining strong evidence in favor of either of the hypotheses when that hypothesis is true is equal to a predetermined level γ. For example, letting "strong" evidence be defined by a likelihood ratio of $k = 8$, then with $\gamma = 0.8$ and $\delta^2/(2\tau^2) = 0.04$, Table 2.3 of Royall (1997, p. 54) yields a total sample size of approximately 230. (Royall's σ^2 equals $2\tau^2$ because we are considering a two-sample problem.) Under this design the probability of observing misleading evidence in favor of the drug is 0.014, and the probability of observing misleading evidence in favor of the placebo is also 0.014. However, with interim monitoring the probability of obtaining misleading evidence will increase. For example, if five equally spaced analyses are conducted (four interim looks and the final analysis) the probability of either type of misleading evidence increases to 0.049; with up to 230 analyses this probability is 0.100 (obtained by simulation), and for an arbitrary number of analyses this probability is approximately 0.111 (Blume, 2002) and is roughly bounded by $1/k = 1/8$ (Robbins, 1970). Thus, there is a price to be paid for the multiple looks at the data with the likelihood ratio paradigm. In addition, there is no practical modification of the design that maintains the interim analyses and retains the small 0.014 probability of observing misleading evidence (likelihood ratio $\geq$8).

What about the stopping boundary itself? If one stops when strong evidence is observed, then a given value k of the likelihood ratio (deemed strong evidence) corresponds to a stopping boundary; we shall refer to this as the "likelihood-ratio boundary". This boundary can be expressed on the same scale as conventional frequentist boundaries, e.g., using Z-values. For the normal mean model it is: stop with strong evidence of hypothesis B if

$$\text{observed } Z\text{-value} > \frac{\delta\sqrt{n}}{2\sqrt{2}\,\tau} + \frac{\sqrt{2}\,\tau \ln(k)}{\delta\sqrt{n}}$$

stop with strong evidence of hypothesis A if

$$\text{observed } Z\text{-value} < \frac{\delta\sqrt{n}}{2\sqrt{2}\,\tau} - \frac{\sqrt{2}\,\tau \ln(k)}{\delta\sqrt{n}}$$

In particular, for the above example the stopping boundary is displayed for five analyses in Figure 1 ($\circ$'s, solid lines).

The likelihood ratio boundaries are one of various possible boundaries. As any stopping boundary is a decision making tool, and since the likelihood ratio approach is concerned exclusively with evaluation of evidence in the data independently of the decision process, it does not provide a strong rationale for requiring a constant likelihood ratio on the boundary. Moreover, the likelihood ratio boundary depends on the alternative hypothesis. This is a problem because although one may be able to specify the minimum clinically interesting treatment effect, this effect will typically be smaller than the effect expected by the investigators. This leads to the obvious question of which alternative hypothesis should be used for defining a likelihood-ratio boundary? One possibility would be to average the likelihood over a prior probability over the alternative region of Δ, and use this weighted average in place of the likelihood in the numerator of the likelihood ratio. This ratio is a special case of what is known as a Bayes factor (Kass and Raftery, 1995); unlike the likelihood ratio, it does require some prior information. Note that if the mass of the "alternative" prior is mostly above the minimum clinically interesting treatment effect, the resulting stopping boundary for positive results will be less conservative in the early looks than the boundary displayed in Figure 1.

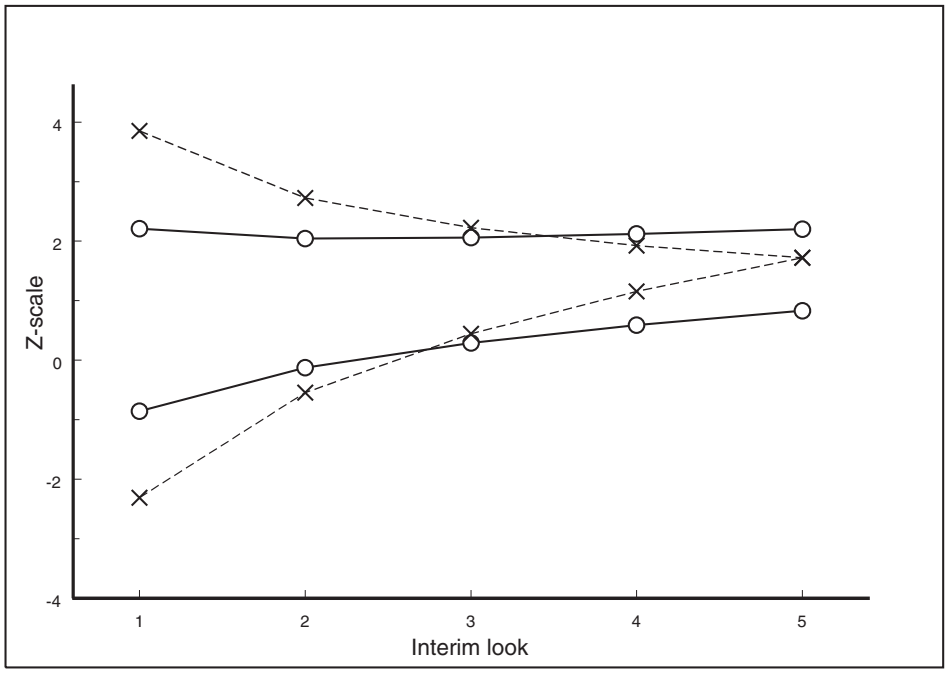

Figure 1 Likelihood ratio interim monitoring boundaries (○, solid line, $k = 8$, $\gamma = 0.8$) and O'Brien–Fleming boundaries (×, dotted line, type 1 error = 0.049) for five equally spaced analyses ($\delta^2/(2\tau^2) = 0.04$, $n = 230$). Stopping can only occur at designated ○ and ×; the connecting lines are for ease of display.

The likelihood-ratio boundary does have the advantage that it suggests stopping with constant level of evidence. In fact, it is closely related to the sequential probability ratio test that was developed to have certain frequentist optimality properties (Wald, 1945). However, other monitoring boundaries like O'Brien–Fleming (O'Brien and Fleming, 1979) or Haybittle–Peto (Haybittle, 1971; Peto et al., 1976) were designed to be very conservative at the interim looks at the data. For example, the implementation of Pampallona and Tsiatis (1994) of the O'Brien–Fleming boundaries are displayed in Figure 1 (× , dotted lines). Conservative boundaries have various properties that make them more desirable than less conservative boundaries like the likelihood-ratio boundaries: (1) they result in only marginal increase in type II error (or increase in sample size) relative to the less conservative boundaries with the same type I error, (2) they prevent early stopping with small sample sizes when data may be too immature to evaluate adequacy of the statistical model and population homogeneity, (3) they help to avoid small trials that often fail to change medical practice, and (4) they minimize problems with trials stopped early for false positive results not being neutralized until many years later when the negative trials come out (Pocock, 1992; Spiegelhalter, Freedman and Parmar, 1994; Ellenberg, Fleming and DeMets, 2002, p. 126).

What about the analysis of a completed trial? A classical frequentist analysis needs to take into account the interim monitoring. In particular, the confidence interval reported should take into account the sample space (defined by the interim monitoring). Although frequently the nominal confidence interval is reported (i.e., ignoring the multiple looks at the data), this is hard to justify from the frequentist point of view except for the fact that the adjustment for interim monitoring will typically be small for reasonably sized trials. For the likelihood-ratio point of view, the likelihood ratios associated with the hypotheses can be reported without any adjustment for the interim monitoring. In theory, one would not even have to report that any interim analyses had been performed.

For additional components of an analysis, the issues are slightly less clear. The frequentist would report an estimator of the treatment effect adjusted for the multiple comparisons. From the likelihood point of view, one could report the maximum likelihood estimate of the treatment effect and a $1/k$ likelihood interval (Royall, p. 26) for the treatment effect (i.e., the set of parameter values Δ for which the likelihood function, standardized by it maximum value, is greater than $1/k$). No adjustment for the interim monitoring would be made. Although no frequentist (repeated-sampling) claims are made, the frequentist properties of the maximum likelihood estimator and the likelihood interval will be affected by the interim monitoring. For example, referring to the example at the beginning of this section, when hypothesis A or B is true, the coverage probability of the 1/8 likelihood interval for Δ will decrease from 95.8% to 92.4% with interim monitoring. This would suggest that it might be useful to report the interim monitoring performed for the benefit of individuals who care about frequentist properties.

With respect to interim monitoring the likelihood ratio quantifies how the prior odds are affected by the evidence in the accumulating data. The posterior odds (the product of the prior odds and the likelihood ratio) are then the current summary of the cumulative beliefs. Without the prior odds the likelihood ratio is thus of limited practical use for making a decision (as is the case for an outcome of a diagnostic test without prevalence information). The decision to stop or continue a trial should be based on the posterior odds. It seems, therefore, that a proper application of the likelihood ratios to interim monitoring can not be separated from Bayesian methods. Bayesian approaches are briefly reviewed in the Discussion.

5 Data Driven Choice of Hypotheses

In randomized clinical trials the primary analysis should be specified in the protocol before the trial begins. This eliminates multiple comparison questions that would arise if the choice of the primary analysis was being driven by the data. An example of how this type of concern can arise is given by a trial of a drug to reduce skeletal-related events for prostate cancer patients with a history of bone metastases (Saad et al., 2002). The original trial design had three treatment arms: a low dose, a high dose, and a placebo. After unacceptable toxicity was seen in patients receiving the high dose, their dose was reduced to the low dose. At the end of the trial the event rates were $92/208 = 44.2\%$ for the placebo, $71/214 = 33.2\%$ for the low dose, and $85/221 = 38.5\%$ for the high $\rightarrow$ low dose, with the placebo versus low dose comparison reported as primary. As was noted (Canil and Tannock, 2002), one would have expected the high $\rightarrow$ low dose results to be at least as good as the low dose results. A theoretical multiple comparisons concern would be if the investigators decided which analysis to report as primary after the data had been seen, e.g., if the high $\rightarrow$ low dose results had been better, then pooling these results with the low dose results. This theoretical concern was not an issue for this trial, however, since the investigators amended the statistical plan of the study before the study was unblinded.

We now consider a hypothetical trial motivated by an example (Berger and Wolpert, 1988, p. 79) to demonstrate the shortcomings of a loose interpretation of likelihood methods. In a clinical trial, patients were randomized between two new cholesterol lowering drugs (D1 and D2) and a placebo control (C). The observed mean drops in cholesterol are given in Table 1. Assume that the drops in

Table 1 Hypothetical data from a clinical trial of two drugs and a placebo control.

	C	D1	D2
Mean $\pm$ $\sigma/\sqrt{n}$	5 ± 6.7	18 ± 6.7	-3 ± 6.7
Sample size (n)	20	20	20

Table 2 Table 1 data with C and D2 pooled to form "Control" treatment.

	"Control"	D1
Mean $\pm$ $\sigma/\sqrt{n}$	1 ± 4.7	18 ± 6.7
Sample size (n)	40	20

cholesterol follow normal distributions with means μ_C, μ_{D1} and μ_{D2}, respectively, with known variance $\sigma^2 = 30^2$. In evaluating drug D1, the investigator might use the following parameterization H_{01}: $\mu_{D1} = \mu_C$ vs. H_{11}: $\mu_{D1} = \mu_C + \Delta$, and report a likelihood ratio of 2 using the known σ^2 (for $\Delta = 20$ as in Section 2). On the other hand, possibly after looking at the data, the investigator may decide that D1 is the main treatment of interest and D2 is worthless with the same effect as the placebo C. He might then test H_{02}: $\mu_C = \mu_{D1} = \mu_{D2}$ vs. H_{12}: $\mu_{D1} = \mu_C + \Delta = \mu_{D2} + \Delta$ and report the likelihood ratio of 8. Moreover, the investigator might present the data as Table 2 (where "Control" refers to pooled C and D2) and not mention that the data were originally collected as Table 1. The investigator believes that the latter approach might be appropriate from the likelihood ratio methods. While we agree that the likelihood ratio of 8 is a summary of the observed evidence for H_{02} vs. H_{12} (D1 vs. "Control"), this type of data driven approach to analysis and reporting of scientific experiments would make interpretation of the results impossible.

The above example is related to various adaptive or two-stage procedures that use the data to select the final analysis model in factorial and crossover trials. There are other scientific areas, e.g. evaluation of DNA evidence in forensic identification, where data driven choice of hypotheses leads to problems with use of the likelihood ratio to measure the strength of evidence of one hypothesis over another (see Meester and Sjerps (2003) and references therein).

6 Discussion

We agree that the likelihood ratio can be used as a measure of statistical evidence comparing two simple hypotheses if one is interested in an abstract measure of evidence independent of making decisions or recommendations, and presumably this is what Blume (2002) had in mind. (*P*-values, although potentially useful for type I error control, are generally thought not to be a good abstract measure of evidence (Casella and Berger, 1987; Berger and Sellke, 1987).) However, clinical trials *are* usually aimed at helping to make decisions or recommendations, in particular, those affecting clinical practice. At the design stage, data are not yet observed and it is necessary to ensure that the design has acceptable characteristics over all possible outcomes. Thus, a frequentist measure, which by definition considers all possible outcomes, would appear to be necessary (Berger, 1985, p. 32). Since frequentist measures must be adjusted for multiple comparisons, this implies that multiple comparisons would need to be considered and accounted for in the design of clinical trials. After the data have been observed, interpretation of the likelihood ratio for making decisions or recommendations seems inseparable from the Bayesian paradigm and prior beliefs, which we discuss below. It should be noted that for questions of goodness of fit, the likelihood ratio may not contain all the relevant information (Lindsey, 1997).

In a Bayesian analysis involving subgroups, one needs to consider the multivariate prior distribution of parameters associated with the treatment effects in the subgroups. If one assumes the parameters are independent, then the analysis of one subgroup will not depend on the observed data in the other subgroups. However, even in this independent prior situation, it has been suggested that one might want to take into account the multiple comparisons by modifying the individual prior distributions to make the null hypotheses more likely and thus keep the prior probability of the global null in accordance with prior belief (Berry, 1990; Westfall, Johnson and Utts, 1997). In particular, one cannot

simply use independent "noninformative" priors as this will lead to unreasonably high posterior probabilities of non-trivial treatment effects when there are many subgroups. An additional consideration is that the priors of the consumers of a subgroup analysis may depend on the number of subgroups considered. For example, our prior belief that the subgroup treatment effect is real in the hypothetical trials described in Section 2 is higher when the subgroup is prespecified by the investigator (and presumably approved by the sponsor or funding agency of the trial). For a prior with non-independent distributions, the data from the other subgroups influence the inference for the subgroup under consideration. This can be done informally, or more formally by assuming a hierarchical prior on the subgroup-associated parameters (Berry 1988; Dixon and Simon, 1991; Berry and Hochberg, 1999). This type of approach typically results in shrinkage of effect sizes for subgroups, which can be viewed as an effect of the multiple comparisons. Bayesian analyses involving multiple variables have similar considerations to subgroup analyses, but with the multivariate prior distribution of parameters being associated with the different variables. An empirical Bayes approach is also possible, in which (for the simplest case) the parameters are assumed to be independent and identically distributed from a (univariate) distribution which is estimated form the data (Morris, 1983; Efron, 2005). In this approach, the analysis of a variable under consideration will again be affected by the data from the other variables.

For Bayesian interim monitoring of a clinical trial, one requires a prior distribution on the treatment effect. This can be used to stop the trial early if, for example, the posterior probability that the treatment effect is "interesting" is high (Berry, 1985; Fayers, Ashby and Parmar, 1997). To make the early stopping more difficult and to take into account the entire range of prior distributions, a "skeptical" ("enthusiastic") prior should be used for stopping early for positive results (futility) (Spiegelhalter et al., 1994). It has been suggested that one might want to modify one's prior based on the type of stopping rule if one thought the choice of stopping rule by the investigator implied some knowledge about the treatment effect (Berger and Wolpert, p. 81). In practice, we do not believe this is a serious concern as long as appropriate skeptical and/or enthusiastic priors are used for monitoring. A Bayesian analysis would appear to require only the range of priors considered and the data at the completion of the trial without information about whether this completion was at an early time.

When the hypotheses are chosen on the basis of the data (as in Section 5), then a Bayesian analysis will require a multivariate prior concerning all the hypotheses. For the example in Table 1, this multivariate prior should reflect the prior beliefs about the relative efficacy of the treatment arms. If the analysis is data-driven, then the hypothesis being considered may not have been thought of before the data were seen. It may be difficult to develop a multivariate "prior" after seeing the data. As a consumer of the analysis, our prior probability that C and D2 have the same treatment effect would be higher if the investigator intended them to be analyzed together then when he decided that after seeing the data.

In summary, at the design stage the investigator should take into account all possible outcomes to ensure that the frequentist operating characteristics of the proposed design are adequate. Thus, with the likelihood ratio approach, multiple comparisons do need to be considered and will affect the design. A proper application of the likelihood ratio to monitoring is virtually inseparable from the Bayesian approach. At the analysis stage, the fact that interim monitoring occurred need not be taken into account for a likelihood or Bayesian analysis. On the other hand, for subgroup analysis, analyses involving multiple variables, or analysis of data-driven hypothesis, investigators need to report the trial data in their entirety so that likelihood ratio results can be interpreted correctly using informal or formal Bayesian analyses. Multiple comparisons will typically affect the conclusions of such Bayesian analyses. Therefore, multiple comparisons issues will need to be considered for the proper interpretation of likelihood ratios. One could argue that the quotes given in the Introduction do not contradict this conclusion, as they do not refer to the interpretation of data but only the "evaluation of evidence" and the "re-examination of accumulating evidence". However, we fear that this distinction may be lost on investigators who would prefer not to pay a price for multiple comparisons.

Acknowledgements The authors would like to thank the referees for their helpful comments.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd ed.). Springer-Verlag, New York.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcibility of P values and evidence (with discussion). *Journal of the American Statistical Association* **82**, 112–139.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. Institute of Mathematical Statistics, Hayward, California.

Berry, D. A. (1985). Interim analyses in clinical trials: Classical vs. Bayesian approaches, *Statistics in Medicine* **4**, 521–526.

Berry, D. A. (1988). Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. in *Bayesian Statistics 3* (Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. M. F., eds.). Oxford University Press, Oxford, U.K.

Berry, D. A. (1990). Subgroup analysis. *Biometrics* **46**, 1227–1230.

Berry, D. A. and Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference* **82**, 215–227.

Blume, J. D. (2002). Tutorial in biostatistics: Likelihood methods for measuring statistical evidence. *Statistics in Medicine* **21**, 2563–2599.

Canil, C. M. and Tannock, I. F. (2002). Should bisphosphonates be used routinely in patients with prostate cancer metastatic to the bone? *Journal of the National Cancer Institute* **94**, 1422–1423.

Casella, G. and Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem (with discussion). *Journal of the American Statistical Association* **82**, 106–139.

Cohen, J. (2003a). AIDS vaccine trial produces disappointment and confusion. Sc*ience* **299**, 1290–1291.

Cohen, J. (2003b). Vaccine results lose significance under scrutiny. *Science* **299**, 1495.

Collins, R., Gray, R., Godwin, J. and Peto, R. (1987). Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews. *Statistics in Medicine* **6**, 245–250.

Dignam, J. J., Bryant, J., Wieand, H. S., Fisher, B., and Wolmark, N. (1998). Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project. *Controlled Clinical Trials* **19**, 575–588.

Dixon, D. O. and Simon, R. (1991). Bayesian Subset Analysis. *Biometrics* **47**, 871–881.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press, London.

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association* **100**, 1–5.

Ellenberg, S. S., Fleming, T. R., and DeMets, D. L. (2002). *Data Monitoring Committees in Clinical Trials*. Wiley, New York.

Fayers, P. M., Ashby, D., and Parmar, M. K. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine* **16**, 1413–1430.

Fleming, T. R., Neaton, J. D., Goldman, A., DeMets, D. L., Launer, C., Korvick, J., Abrams, D., and the Terry Beirn Community Programs for Clinical Research on AIDS (1995). Insights from monitoring CPCRA didanosine/zalcitabine trial. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* **10** (Suppl. 2), S9–S18.

Good, I. J. (1985). Weight of evidence: a brief survey. In *Bayesian Statistics 2* (Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. M. F., eds.). Elsevier, Amsterdam, pp. 249–270.

Hacking, J. (1965). *Logic of Statistical Evidence*. Cambridge University Press, New York.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44**, 793–797.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford University Press, Oxford, U.K.

Kass, R. E. and Raftery A. E. (1995). Bayes Factor. *Journal of the American Statistical* Association **90**, 773–795.

Lindsey, J. K. (1997). Stopping rules and the likelihood function. *Journal of Statistical Planning and Inference* **59**, 166–177.

Meester, R. and Sjerps, M. (2003). The evidential value in the DNA database search controversy and the two-stain problem. *Biometrics* **59**, 727–732.

Miller, R. G. (1981). *Simultaneous Statistical Evidence* 2nd ed. Springer-Verlag, New York.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47–55.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pampallona, S. and Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* **42**, 19–35.

Peto, R. (1996). Five years of tamoxifen – or more? *Journal of the National Cancer Institute* **88**, 1791–1793.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient: introduction. *British Journal of Cancer* **34**, 585–612.

Pocock, S. J. (1992). When to stop a clinical trial. *British Medical Journal* **305**, 235–240.

Robbins H. (1970) Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics* **41**, 1397–1409.

Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, London.

Royall, R. M. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association* **95**, 760–768.

Saad, F., Gleason, D. M., Murray, R., Tchekmedyian, S., Venner, P., Lacombe, L., Chin, J. L., Vinholes, J. J., Goas, J. A., and Chen, B. (2002). A randomized, placebo-controlled trial of zoledronic acid in patients with hormone-refractory metastatic prostate carcinoma. *Journal of the National Cancer Institute* **94**, 1458–1468.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society, Ser. A* **157**, 357–416.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**, 117–186.

Westfall, P. H., Johnson W. O., and Utts J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika* **84**, 419–427.

