STATISTICS IN MEDICINE
Statist. Med. **17**, 1779–1797 (1998)

MULTI-CENTRE TRIAL ANALYSIS REVISITED

A. LAWRENCE GOULD*

Merck Research Laboratories, West Point, PA 19486, U.S.A.

SUMMARY

Analyses of multi-centre trials must consider the effects of the individual centres and the possibility of non-constancy of treatment effect differences among centres. This usually means an ANOVA with terms for centres, treatments, and centre × treatment interactions in practice, at least in the U.S.A. Empirical and conventional Bayes methods provide attractive alternatives to conventional ANOVAs for analysing and reporting the findings from multi-centre trials and do not require more restrictive assumptions than the ANOVA approach. These approaches require regarding the centre effects as random instead of fixed, a view which often will reasonably describe outcomes of clinical trials in spite of the fact that the individual centres certainly do not comprise a random sample of all possible centres. The components of these approaches are well understood and have been employed in related applications such as meta-analysis. Combining them in a way that makes their application to routine muti-centre trial analysis relatively straightforward does not appear to have been described previously, and is what forms the topic of this paper. The empirical Bayes approach leads to useful graphical displays, including one with the data superimposed on probability contours of the joint distribution of the individual centre means and standard deviations, which provides a handy way to identify possible outliers. Covariates can be incorporated without difficulty. The Bayes approach, implemented with Gibbs sampling, provides a convenient way to construct posterior and predictive distributions for a variety of useful statistics. We compare the result of empirical and conventional Bayes analyses with the result of fixed and mixed model ANOVAs applied to data from a multi-centre trial.
© 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Analyses of multi-centre trials must consider the effects of the individual centres and the possibility of non-constancy of treatment effect differences among centres. Some 10 years ago, Fleiss wrote 'The most challenging questions in the analysis of the data from a multicenter trial are how to carry out the analysis when there is treatment-by-center interaction, and, prior to that, how to ascertain whether such interaction exists.'[1] Then, and today, the conventional analysis, at least in the U.S.A., is a fixed effects ANOVA with terms for centres, treatments, and centre × treatment interactions. Whether this represents the best way to approach the problem remains to be seen. However, the need to assure consistency of inferences about treatment effects across centres and to provide adequate supporting documentation transcends methodologic details, and

* Correspondence to: A. Lawrence Gould, Merck Research Laboratories, West Point, PA 19486, U.S.A.

CCC 0277–6715/98/161779–19$17.50
© 1998 John Wiley & Sons, Ltd.

can be accomplished in various ways. Empirical and conventional Bayes approaches present alternatives that address these needs and provide useful additional information.

The definition of 'interaction' is a central issue. If centres are viewed as fixed effects,[1] then interaction can be defined as variation in the true treatment fixed effects over levels of the true centre fixed effects. If the *observed* variation in the treatment effects across centres exceeds what can be explained by sampling error alone, then either the model expressing what chance alone would predict is wrong, or some of the centres possess attributes affecting the actions of the treatments (true interaction) and a serious attempt must be made to identify these attributes. Which of these situations applies in any instance is a judgement call.

As a useful operational alternative, unless the finding from one or a few centres differs substantially from the findings for the remaining centres, it probably is reasonable to regard the variation as random, especially if there are many centres and 'outlying' centres do not otherwise differ materially from the other centres. This can be a convenient approximation even though it certainly is true that centres are not randomly sampled from some plausible population of centres. If one rejects the extreme view that centres and centre × treatment interaction should be regarded as fixed effects, then some way to accommodate excess variability among centres is needed. This can be accomplished in various ways. The 'pushback' method by Ciminera *et al.*[2,3] provides an imaginative approach. Alternatively, one can employ meta-analytic methods such as mixed model ANOVA (or, more generally, generalized linear model (GLM) or mixed model (GLMM) techniques), Empirical or full Bayes approaches,[4] random effect likelihood models,[5–7] multi-level modelling,[8] and hierarchical Bayes linear models[9] because the analysis of multi-centre trial outcomes is a meta-analysis in the sense of summarizing the information each centre provides about the treatment differences.

This paper focuses on Bayes and empirical Bayes methods as alternatives to the conventional fixed or mixed model ANOVA for analysing data from multi-centre trials. They are compared in the context of the analysis of a large multi-centre trial comparing finasteride against placebo in the treatment of benign prostatic hyperplasia.[10] Empirical and conventional Bayes methods provide attractive alternatives for analysing and reporting the findings from multi-centre trials and do not require more restrictive assumptions than the ANOVA approach. They can be applied to non-normally distributed data such as count data or binary data, with appropriate changes in computational details.

2. MULTI-CENTRE TRIAL

The efficacy and tolerability of finasteride, a drug for treating benign prostatic hyperplasia, were studied in 25 centres in the U.S. and 5 centres in Canada in a multi-centre trial of approximately 900 men with symptoms of urinary obstruction, an enlarged prostate gland on digital rectal examination, and maximal urinary flow rates of less than 15 ml/second.[10] The trial entrants were assigned at random to receive finasteride 5 mg, finasteride 1mg, or a matching placebo once a day. Table I provides summary statistics for change from baseline in total symptom score for each centre included in this analysis. The total symptom score is the sum of the responses to nine questions about symptoms pertaining to various aspects of impaired urinary ability. Each symptom was scored on a 5-point scale, with 0 (symptom absent) to 4 (symptom severe), leading to a range of total scores between 0 and 36. Central limit considerations suggest that it is reasonable to assume that the total symptom scores are at least approximately normally distributed. One especially sparse centre, contributing a total of three patients, was omitted to

© 1998 John Wiley & Sons, Ltd.

Table I. Summary of findings from multi-centre trial of finasteride. The responses are changes in total symptom score from baseline

Centre	N	Placebo		Finasteride 1 mg			Finasteride 5 mg		
		Mean	Standard deviation	N	Mean	Standard deviation	N	Mean	Standard deviation
1	7	0·43	4·58	7	− 5·43	5·53	8	− 2·63	3·38
2	11	0·10	4·21	11	− 2·59	3·95	12	− 2·21	4·14
3	6	2·58	4·80	6	− 3·94	4·25	7	1·29	7·39
4	10	− 2·30	3·86	10	− 1·23	5·17	10	− 1·40	2·27
5	10	2·08	6·46	10	− 6·70	7·45	10	− 5·13	3·91
6	6	1·13	3·24	5	3·40	8·17	5	− 1·59	3·19
7	5	1·20	7·85	6	− 3·67	4·89	5	− 1·40	2·61
8	12	− 1·21	2·66	13	0·18	3·81	12	− 4·08	6·32
9	8	1·13	5·28	8	− 2·19	5·17	9	− 1·96	5·84
10	9	− 0·11	3·62	10	− 2·00	5·35	10	0·60	3·53
11	15	− 4·37	6·12	14	− 2·68	5·34	15	− 2·14	4·27
12	8	− 1·06	5·27	9	0·44	4·39	9	− 2·03	5·76
13	12	− 0·08	3·32	12	− 4·60	6·16	11	− 6·22	5·33
14	9	0·00	5·20	9	− 0·25	8·23	7	− 3·29	5·12
15	6	1·83	5·85	7	− 1·23	4·33	6	− 1·00	2·61
16	14	− 4·21	7·53	14	− 2·10	5·78	12	− 5·75	5·63
17	13	0·76	3·82	13	0·55	2·53	13	− 0·63	5·41
18	15	− 1·05	4·54	13	2·54	4·16	14	− 2·80	2·89
19	15	2·07	4·88	15	− 1·67	4·95	15	− 3·43	4·71
20	11	− 1·46	5·48	10	− 1·99	5·63	10	− 6·77	5·19
21	5	0·80	4·21	5	− 3·35	4·73	5	− 0·23	4·14
22	11	− 2·92	5·42	10	− 1·22	5·95	11	− 4·45	6·65
23	9	− 3·37	4·73	9	− 1·38	4·17	7	0·57	2·70
24	12	− 1·92	2·91	12	− 0·66	3·55	12	− 2·39	2·27
25	9	− 3·89	4·76	9	− 3·22	5·54	8	− 1·23	4·91
26	15	− 3·48	5·98	15	− 2·13	3·25	14	− 3·71	5·30
27	11	− 1·91	6·49	12	− 1·33	4·40	11	− 1·52	4·68
28	10	− 2·66	3·80	10	− 1·29	3·18	10	− 4·70	3·43
29	13	− 0·77	4·73	13	− 2·31	3·88	13	− 0·47	4·95

preserve numerical stability. We focus on the comparison of the 1 mg dosage with placebo, but occasionally consider briefly the comparisons with the 5 mg dosage.

3. ANALYSES OF VARIANCE

3.1. Fixed effect ANOVA

A fixed effect analysis of variance on the data summarized in Table I yields the results given in Table II. The calculations were performed using the SAS GLM procedure.[11] The model used for this analysis is the conventional two-way factorial model with interaction:

$$y_{ijk} = \mu + b_i + t_j + (bt)_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, B; j = 1, 2; k = 1, \ldots n_{ij} \tag{1}$$

© 1998 John Wiley & Sons, Ltd.

Table II. ANOVA tables, estimates and 95 per cent confidence intervals for fixed effect model (** = $p < 0{\cdot}01$, *** = $p < 0{\cdot}001$; all tests are at a nominal 5 per cent level)

Source	All groups		Finasteride 1 mg and Placebo		Finasteride 5 mg and Placebo	
	d.f.	MS	d.f.	MS	d.f.	MS
Centre	28	47·10**	28	36·58	28	48·55**
Treatment	2	191·03***	1	64·22	1	395·41***
Centre × trt	56	38·11**	28	46·58**	28	33·81
Error	798	24·16	536	25·16	530	23·77
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Finasteride 1 mg - Placebo	− 0·66	− 1·63, 0·31	− 0·66	− 1·47, 0·15	–	
Finasteride 5 mg - Placebo	− 1·58	− 2·55, −0·60	–		− 1·58	− 2·37, − 0·79

where y_{ijk} denotes the observation on the kth patient on treatment j in centre i, μ denotes the overall mean, b_i denotes the (fixed) effect of the ith centre, t_j denotes the effect of the jth treatment, $(bt)_{ij}$ denotes the (fixed) interaction effect, and ε_{ijk} denotes a residual error distributed with zero mean and variance σ_ε^2. The centre × treatment interaction term is highly significant. This interaction appears to be due to the variation among centres in the differences between the responses to finasteride 1 mg and placebo. The interaction contribution is not significant when only the finasteride 5 mg and placebo groups are considered, but there is a significant treatment effect. Interaction test findings should be used cautiously to choose among models because tests for interaction have low power; a non-significant finding does not necessarily mean that there is no interaction. It is not clear whether these findings should be interpreted as true inconsistency of treatment effect across centres or merely substantial random variation among centres. Tests can be carried out to determine if the 'interaction' is a matter of direction or of degree,[12] but these tests do not resolve the interpretation issue.

3.1. Mixed Model ANOVA

A mixed model, or random effects, ANOVA assumes that the centre and interaction effects are random, that is, that the observable treatment differences can vary among centres, for example, because of random differences in disease severity or manifestation. Table III provides the results of applying a mixed model analysis to the data summarized in Table I. The model is similar to (1):

$$y_{ijk} = \mu + \beta_i + t_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, \ldots, B; j = 1, 2; k = 1, \ldots, n_{ij} \tag{2}$$

but with Greek instead of Roman letters used to identify the random factors (except for μ, which is fixed). We employ the usual assumptions of mutual independence and normality, $\beta_i \sim \mathrm{N}(0, \sigma_\beta^2)$, $\gamma_{ij} \sim \mathrm{N}(0, \sigma_\gamma^2)$, and $\varepsilon \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. The calculations including all of the treatment groups and each finasteride group separately were carried out using the SAS MIXED procedure with REML (REstricted Maximum Likelihood) estimation.[11] REML estimates are similar to ML estimates except that the downward bias of ML estimates is eliminated by basing the estimates of the variance components only on error contrasts of the data.[13] Solutions of REML equations for

© 1998 John Wiley & Sons, Ltd.

Table III. Estimates and 95 per cent confidence intervals from mixed effects ANOVA calculations (** = $p < 0{\cdot}01$, *** = $p < 0{\cdot}001$; all tests are at a nominal 5 per cent level)

	Finasteride 1 mg - Placebo			Finasteride 5 mg - Placebo			σ_ε^2	σ_γ^2
	Estimate	Confidence	Interval	Estimate	Confidence	Interval		
All groups (SAS)	− 0·77	− 1·80,	0·25	− 1·61**	− 2·64,	− 0·58	24·16	1·36
All groups (S-plus)	− 0·75	− 1·73,	0·24	− 1·61**	− 2·60,	− 0·63	23·64	1·28
Finasteride 1 mg and Placebo	− 0·78	− 1·88,	0·31	–			25·14	1·60
Finasteride 5 mg and Placebo	–			− 1·62**	− 2·62,	− 0·61	23·70	1·11

fixed effects and variance components are the same as ANOVA estimates from balanced designs obtained by equating ANOVA sums of squares to their expectations.[14] Robust variance component estimates were calculated using the Winsorized REML method provided with the S-plus VARCOMP procedure.[15] The robust estimates here provide slightly narrower confidence intervals for the treatment differences than the estimates from SAS.

These findings are consistent with those in Table II. The confidence intervals for the comparisons between finasteride and placebo using the random effects model are wider than with the fixed effects model because of the centre × treatment random effect. This analysis, while more insightful than the fixed effect model analysis, still needs to be supplemented to identify centres whose estimated treatment effects differ by more than what chance would predict from the estimates provided by the general body of centres. Further examination of these centres is needed to explain the excess variation.

4. EMPIRICAL BAYES ANALYSIS

4.1. Common features of Bayes and Empirical Bayes analyses

The analyses described here are based on empirical Bayes[16] and parametric empirical Bayes[17,18] principles. The ith centre provides n_{fi} observations on finasteride and n_{pi} observations on placebo drawn from $N(\mu_{fi}, \sigma_i^2)$, and $N(\mu_{pi}, \sigma_i^2)$ distributions. The difference between the group sample means $d_i = \bar{x}_{fi} - \bar{x}_{pi}$ for the ith centre estimates the true difference $\delta_i = \mu_{fi} - \mu_{pi}$ between the mean changes from baseline in the finasteride and placebo groups for that centre. Combining the sample variances s_{fi}^2, s_{pi}^2 for each group in the ith centre provides an estimate $s_i^2 = (m_{fi}s_{fi}^2 + m_{pi}s_{pi}^2)/m_i$ of the common within-group variance $\sigma_i^2 (= \sigma_\varepsilon^2)$, where $m_{fi}(m_{pi})$ denotes the degrees of freedom associated with $s_{fi}^2(s_{pi}^2)$; $m_{fi} = n_{fi} - 1$ and $m_{pi} = n_{pi} - 1$ in this example, $m_i = m_{fi} + m_{pi}$. The sample likelihood corresponding to the ith centre can be expressed as the product of a chi-square density with m_i degrees of freedom (d.f.) and a standard normal density:

$$f(d_i, s_i^2; n_{fi}, n_{pi}, \delta_i, \sigma_i^2) = \frac{m_i}{\sigma_i^2} f_{\chi^2}\left(\frac{m_i s_i^2}{\sigma_i^2}; m_i\right) \times \frac{1}{\sigma_i}\sqrt{\tilde{n}_i}\,\phi\left(\frac{(d_i - \delta_i)}{\sigma_i}\sqrt{\tilde{n}_i}\right),$$

$$\tilde{n}_i = \frac{n_{fi}n_{pi}}{n_{fi} + n_{pi}}. \tag{3}$$

Expression (3) reflects only sampling error within the ith centre. However, the variation observed among the outcomes of a collection of centres can exceed what can be accounted for by sampling

© 1998 John Wiley & Sons, Ltd.

error alone. The values of δ_i and σ_i^2 may vary among centres, reflecting possible differences in their populations, severity of disease etc. This extra-sampling variation can be addressed by postulating the existence of a prior distribution for δ_i and σ_i^2 among centres from which values are drawn at random for any particular centre. A convenient functional form for this prior distribution[19] is similar to (3):

$$g(\delta, \tau; \Delta, \zeta, \eta\omega) = \omega\zeta f_{\chi^2}(\omega\zeta\tau; \omega) \times \sqrt{(\eta\tau)}\phi((\delta - \Delta)\sqrt{(\eta\tau)}) \tag{4}$$

where $\tau = 1/\sigma^2$. The joint density of the outcomes $(d_i, s_i^2, \delta_i, \tau_i)$ is the product of (3) and (4). Integrating this product with respect to δ_i and τ_i yields the marginal density of d_i and s_i^2 that reflects both sampling variation within centres and extra-sampling variation among centres. The joint marginal density of $u_i = (d_i - \Delta) \times \sqrt{\{\tilde{n}_i\eta(m_i + \omega)\ (1 - v_i)/\omega\zeta\ (\tilde{n}_i + \eta)\}}$ and $v_i = m_i s_i^2/(m_i s_i^2 + \omega\zeta)$ has a particularly simple expression:

$$h_i(u_i, v_i; \Delta, \zeta, \eta\omega) = f_\beta\left(v_i; \frac{m_i}{2}, \frac{\omega}{2}\right) \times f_t(u_i; m_i + \omega) \tag{5}$$

where i indexes the centres in the collection, $f_\beta(.; a, b)$ denotes a central beta density with (a, b) d.f. and $f_t(.; k)$ denotes a central t density with k d.f. The marginal likelihood for the trial is the product of expressions (5) calculated for each centre in the trial. The essential computational distinction between the empirical and conventional Bayes approaches occurs at this point. The empirical Bayes approach maximizes the marginal likelihood (or its logarithm) with respect to the values of Δ, ζ, η and ω, yielding estimates $\hat{\Delta}$, $\hat{\zeta}$, $\hat{\eta}$ and $\hat{\omega}$. These estimates, or monotonic transformations of them, are assumed to have a multivariate normal distribution. The Bayes approach, described in more detail below, does not make this assumption. Substituting $\hat{\Delta}$, $\hat{\zeta}$, $\hat{\eta}$ and $\hat{\omega}$ into (4) provides the *joint predictive density* of δ and τ, which describes the variation in the centre-specific parameters δ_i and τ_i (or σ_i^2) that the current evidence suggests should be expected in similar future trials.

4.2. Variability of estimators

The variability of the maximum likelihood estimators can be determined in various ways. The familiar Wald variance estimators are obtained by substituting the parameter estimates into the expressions for the second derivatives for the log-likelihoods. These variance estimators can underestimate the true variability if the likelihood surface is not well behaved in a neighbourhood of the maximum. The profile likelihood can be used to provide confidence intervals with better coverage properties.[20]

4.3. Checking model fit

The assumptions underlying the analysis can be checked by a straightforward goodness-of-fit test.[21] The CDF of $Z = F(X,Y)$, the random variable defined as the CDF of (X,Y) evaluated for X and Y considered as random variables, can be written as $G_z(z) = z(1 - \ln(z))$. If the model is correct, then (5) should describe the marginal distribution of the study outcomes, and the calculated z values should have CDF $G_Z(z)$, so that a plot of the $(z, G_Z(z))$ pairs should lie on a 45 degree line through the origin. If they do not (as measured by the GOF test), then the model is not correct. If they do, the model may be correct[22] (and, in practice, probably is).

© 1998 John Wiley & Sons, Ltd.

Table IV. Empirical Bayes results (finasteride 1 mg and placebo)

		Δ	$\xi = \ln(\zeta)$	$\ln(\eta)$	$\ln(\omega)$
Estimate		0·639	3·161	1·999	3·291
Standard error		0·539	0·084	0·685	0·601
95% CI (Wald)		− 1·75, 0·47	2·99, 3·33	0·59, 3·41	2·05, 4·53
95% CI (profile)		− 1·83, 0·48	2·97, 3·33	0·85, 9·75	2·25, 5·46
Correlations	Δ	1	− 0·007	0·119	− 0·102
	$\xi = 1n(\zeta)$		1	0·151	0·248
	$\ln(\eta)$			1	− 0·056

4.4. Computational results

Table IV and Figure 1 summarize the empirical Bayes calculations applied to the data for the finasteride 1 mg and placebo groups in Table I. The calculations were carried out using the maximum likelihood routines in the GAUSS™ [23] programming system. The estimates of ζ, η and ω are, respectively, 23·60, 7·38 and 26·88. The Wald and profile confidence intervals are quite close for all of the parameters except $\ln(\eta)$. This suggests that inferences based on a multinormal model for Δ and $\xi = \ln(\zeta)$ are likely to be realistic. The left side of Figure 1 illustrates the goodness-of-fit test. The findings from the test suggest that the model describes the observed outcomes reasonably well. The right side superimposes a plot of the joint conditional (on the maximum likelihood estimate values) predictive distribution of $(\sigma, \delta - \hat{\Delta})$ and a scatter plot of $(s, d - \hat{\Delta})$, where $\hat{\Delta}$ is the maximum likelihood estimate of Δ. The joint conditional predictive density is given by (4) with the maximum likelihood estimates substituted for the parameters. There is only one apparently outlying point (centre 5). Rerunning the fixed effect ANOVA without this centre caused the interaction to become non-significant in the analysis including only the finasteride 1 mg and placebo groups, and only marginally significant ($p = 0{\cdot}037$) in the analysis including all of the data. For comparison with the results from the fixed and mixed ANOVAs, the empirical Bayes estimate and confidence interval for the true finasteride-placebo difference is obtained from the statistics for Δ: $-0{\cdot}64$ ($-1{\cdot}8$, 0·5) and the estimate and confidence interval for $\sigma^2 = \sigma_\varepsilon^2$ is obtained from the statistics for $\zeta = \exp(\xi)$: 23·6 (19·5, 27·9).

The variances of δ and τ obtained from (4) underestimate the true variances by an amount that depends on the variability of $\hat{\Delta}$, $\hat{\zeta}$, $\hat{\eta}$ and $\hat{\omega}$.[24] A better estimate of the true variance is provided by the expression $V(Y) = E_x(V(Y|X)) + V_x(E(Y|X))$. The conditional distribution of $u = (\delta - \Delta)\sqrt{\eta/\zeta}$ given $\hat{\Delta}$, $\hat{\zeta}$, $\hat{\eta}$ and $\hat{\omega}$ is a central t with $\hat{\omega}$ degrees of freedom from (5), so that $V(\delta|\hat{\Delta}, \hat{\zeta}, \hat{\eta}, \hat{\omega}) = (\hat{\zeta}\hat{\omega}/\hat{\eta}(\hat{\omega} - 2)$. If v_Δ^2 denotes the asymptotic variance of $\hat{\Delta}$, then the unconditional variance of δ is approximately

$$V(\delta) = E\{\hat{\zeta}\hat{\omega}/\hat{\eta}(\hat{\omega} - 2)\} + v_\Delta^2.$$

A further approximation can be obtained by replacing the expectation with the estimated quantities. Table IV provides the quantities needed to estimate the unconditional variance of δ, $V(\delta) = (23{\cdot}6)\,(26{\cdot}88)/(7{\cdot}38)\,(24{\cdot}88) + 0{\cdot}539^2 = 3{\cdot}46 + 0{\cdot}29 = 3{\cdot}75$, which only slightly exceeds the conditional variance = 3·46. The error induced by the maximum likelihood estimation process is negligible here, so the conditional calculations are sufficient. This will not always be true,

© 1998 John Wiley & Sons, Ltd.

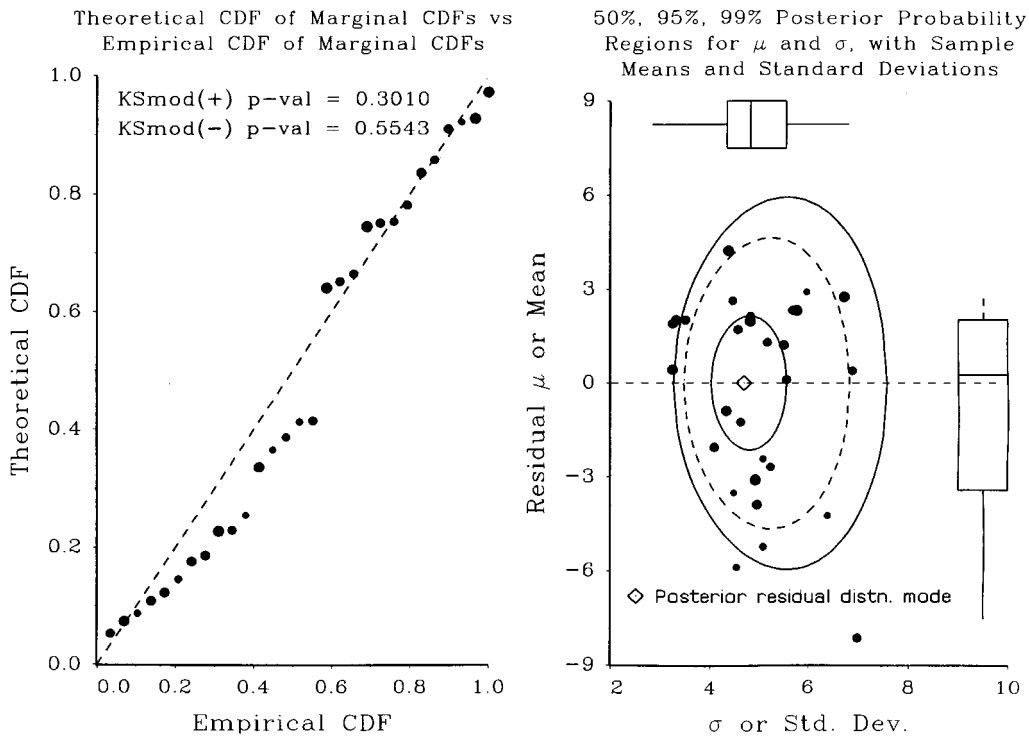

Figure 1. Goodness-of-fit test, and scatter and box plots of observed data superimposed on contours of conditional predictive distribution of δ and σ

however, especially when relatively few centres are included so that the maximum likelihood estimates of the parameters do not have great precision.

5. BAYES ANALYSIS

A Bayesian analysis of these data can proceed in various ways, including Bayesian hypothesis testing and model selection.[25,26] The empirical Bayes and Bayes formulations differ essentially in how the parameters of the prior distribution are handled in the analysis. The empirical Bayes approach constructs the marginal likelihood of the data as a function of the prior distribution parameters, estimates these parameters by maximizing the marginal likelihood, and bases inferences about the prior parameters on the assumption that the estimates have the asymptotic multivariate normal joint distribution characteristic of well-behaved maximum likelihood estimators. The Bayesian approach used here postulates that the parameters of the prior distribution of the centre-specific parameters $\delta_i, \tau_i, i = 1, \ldots, B$ also have a prior distribution which, for convenience, we take to be vague, that is, proper but essentially uninformative. Inferences about these parameters are based on their posterior distribution, which does not have to be multivariate normal.

We consider here two simple Bayesian analyses, one using a simple independence model, and one using the model employed in the empirical Bayes analysis. The simple independence model

© 1998 John Wiley & Sons, Ltd.

replaces the product $\eta\tau$ in (4) with the quantity ψ:

$$g^*(\delta, \tau; \Delta, \zeta, \omega, \psi) = \omega\zeta f_{\chi^2}(\omega\zeta\tau; \omega) \times \sqrt{\psi}\phi((\delta - \Delta)\sqrt{\psi}) \tag{6}$$

δ and τ have independent prior distributions under (6). δ and τ are uncorrelated *a priori*, but not independent, under (4). Expression (6) is arguably a more 'intuitive' expression for the prior, but it is not a natural conjugate prior like (4) is.

Let θ denote collectively the parameters for either formulation [$\theta = (\Delta, \zeta, \omega, \psi)$ or $(\Delta, \zeta, \omega, \eta)$]. The posterior distribution of θ is obtained by dividing the marginal density of the data into the joint density of the data and parameters

$$\prod_{i=1}^{B} f_i(d_i, s_i^2; \delta_i, \tau_i)\ g(\delta_i, \tau_i; \theta)\ p_0(\theta) \tag{7}$$

where f_i and g are given by (3) and (4) or (6). The marginal density of the data is the integral of (7) with respect to parameters, and can be written as

$$\int_\theta \sum_{i=1}^{B} h(u_i, v_i; \theta)\ p_0(\theta)\, \mathrm{d}\theta$$

where h is given by (5) and u_i, v_i are functions of the elements of θ. This integral has no closed expression, so the posterior distribution of θ must be obtained numerically rather than analytically. The posterior distribution can be obtained numerically via Markov chain Monte Carlo (MCMC) techniques such as Gibbs sampling.[27,28]

The MCMC approach generates samples from the distributions of the quantities of interest; kernel density estimation procedures use the samples to describe the distributions. Not just posterior distributions of parameters can be generated in this way. We also can obtain the distribution of any inferentially interesting quantity. The computational aspects of MCMC calculations are not trivial, and readers not familiar with the technical details are advised to use specialized software for this purpose. The BUGS software[29] available from Cambridge University provides a convenient way to do the computations.

We demonstrate here how the BUGS software can be used to analyse date from multi-centre trials. We also illustrate the use of additional display and diagnostic methods that can help provide useful insights. The technique is the same as that used in meta-analysis, since the analysis of multi-centre trial outcomes is a kind of meta-analysis in the sense of summarizing the information each centre provides about the treatment differences.

5.1. Model and setup for MCMC calculations

The primary objectives are inferences about (a) the values of the differences between the effects of finasteride in either dosage against placebo, and (b) the consistency of these difference values across studies. Additional inferences easily can be addressed.

Each centre provides an estimate d of the mean finasteride-placebo difference for that centre (δ), and also an estimate s^2 of the centre-specific within-group variance $\sigma_W^2 = 1/\tau_W$, assumed to be the same for each treatment group in the centre; $\sigma_W^2 = \sigma_\varepsilon^2$ in the ANOVAs. Figure 2 summarizes the dependency relations for the two models using the graphical scheme described in the BUGS manuals. We assume that the centre-specific within-group variances arise from an inverse gamma

© 1998 John Wiley & Sons, Ltd.

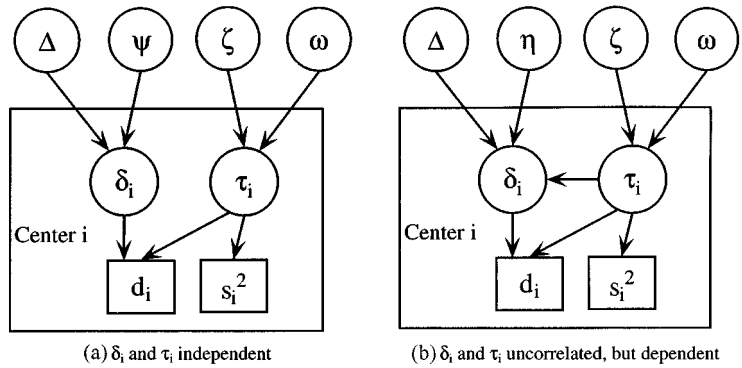

Figure 2. Relationships among the random quantities in the multi-centre models: (*a*) simple independence model; (*b*) model used for empirical Bayes analysis

distribution in both formulations. The centre-specific mean differences are assumed to arise from a normal distribution with mean Δ and variance $\sigma_B^2 = 1/\tau_B$. This quantity is explicit in the independence model and implicit in the model using the natural conjugate prior (4). We use vague, essentially uninformative, priors for simplicity, avoidance of controversy,[30] and because there is no obvious or natural prior in this application. The key portions of the BUGS commands file that specify the calculations to be done for the independence model are included in the Appendix.

5.2. Computational results

Table V summarizes the result of the Bayesian analysis calculations for both priors using the Gibbs sampling scheme implemented in the BUGS package. The calculations for the independence model used an initial burn-in of 2000 replicates, followed by 25,000 replicates; only every 50th replicate was used, so that 500 iterates were available for parameter estimation parameters and most diagnostics. An initial burn-in of 2000 replicates was followed by 50,000 replicates for the model using the natural conjugate prior; every 500th was used. The replicates were thinned to minimize the autocorrelations among the simulated values. The autocorrelations among the successive draws from the posterior distributions of the parameters of the simple model were:

Thinning Interval	$\Delta_{\text{Fin5-Pbo}}$	$\Delta_{\text{Fin1-Pbo}}$	σ_B	σ_W	$\delta_{\text{Fin1-Pbo}}$	$\delta_{\text{Fin5-Pbo}}$
1 (no thinning)	0·69	0·65	0·93	0·06	0·08	0·06
50	0·13	0·06	0·40	0·01	− 0·02	0·03
100	0·05	0·10	0·39	0·03	− 0·04	− 0·03

Thinning to every 50th replicate appeared to reduce the autocorrelations of all the iterates satisfactorily except for σ_B, for which a thinning interval of between 500 and 1000 would be needed. Additional runs of 2000 replicates used starting values consisting of all combinations of $(\Delta_{\text{Fin1-Pbo}}, \Delta_{\text{Fin5-Pbo}}) = (-10, 10)$ or $(10, -10)$, $\sigma_B^{-2} = 1$ or 10, $\sigma_W^2 = 1$ or 20. The statistics of the sets of simulated values from the posterior distributions of $\Delta_{\text{Fin1-Pbo}}$, $\Delta_{\text{Fin5-Pbo}}$, σ_B, σ_W, $\delta_{\text{Fin1-Pbo}}$, and

© 1998 John Wiley & Sons, Ltd.

Table V. Summary of Bayesian analysis calculations: characteristics of posterior distributions

Variable	Mean	Standard deviation	Quantiles of the posterior distribution				
			2·5%	25%	50%	75%	97·5%
(*a*) *Independence prior*							
$\Delta_{F1\text{-}Pbo}$	− 0·639	0·526	− 1·72	− 0·972	− 0·654	− 0·272	− 0·307
$\Delta_{F5\text{-}Pbo}$	− 1·58	0·502	− 2·61	− 1·88	− 1·57	− 1·28	− 0·538
σ_B	1·47	0·552	0·086	1·18	1·54	1·86	2·41
σ_W	4·88	0·549	4·02	4·52	4·86	5·15	6·19
$\delta_{F1\text{-}Pbo}$	− 0·700	1·66	− 4·24	− 1·65	− 0·622	0·264	2·49
$\delta_{F5\text{-}Pbo}$	− 1·58	1·63	− 4·63	− 2·51	− 1·53	− 0·649	1·55
(*b*) *Natural conjugate prior* (*empirical Bayes model*)							
$\Delta_{F1\text{-}Pbo}$	− 0·668	0·487	− 1·64	− 0·971	− 0·709	− 0·351	0·334
$\Delta_{F5\text{-}Pbo}$	− 1·530	0·520	− 2·60	− 1·85	− 1·55	− 1·23	− 0·524
σ_B	1·41	0·671	< 0·01	1·08	1·48	1·88	2·56
σ_W	4·88	0·544	4·00	4·53	4·81	5·16	6·08
$\delta_{F1\text{-}Pbo}$	− 0·686	1·64	− 4·06	− 1·57	− 0·765	0·187	2·76
$\delta_{F5\text{-}Pbo}$	− 1·58	1·67	− 5·08	− 2·52	− 1·56	− 0·683	1·77

$\delta_{Fin5\text{-}Pbo}$ were identical for all of the runs. This suggests that the findings presented in Table V did not depend on the choice of starting values. Section 5.4 discusses briefly the result of applying some of the diagnostic procedures supplied with the BUGS package.

The first rows of each part of Table V summarize the 'true' differences between the 1 mg and 5 mg dosages of finasteride and placebo with respect to change in total symptom score from baseline. As before, negative changes imply improvement, so that the patients on finasteride improved more than the patients on placebo did. The findings for the finasteride 1 mg-placebo comparison are quite similar to those obtained using empirical Bayes approach (Table IV).

The middle rows of each section of Table V provide information about the variation within and between studies. The within-group variances do not vary much across studies, and this is reflected in the narrow confidence intervals for σ_w. The estimate of 4·88 from this analysis is very close to the estimate of 4·86 ($=\sqrt{23{\cdot}60}$ from the empirical Bayes analysis and 4·92 ($=\sqrt{24{\cdot}16}$) from the mixed model ANOVA. The variance of the study-specific mean differences across studies is quite appreciable. This means that although there may be a real treatment effect, the variations among studies, for example, with respect to disease severity (degree of prostatic hyperplasia) may be large enough so that in some studies the study-specific mean difference that is estimated by the sample static d in fact may not be negative.

The bottom two rows of each part of Table V provide statistics for the predictive distributions of the study-specific treatment differences. The variances of the predictive distributions are less than the values obtained from the empirical Bayes analysis (1·93). It is not surprising that the predictive distribution for the finasteride 1 mg-placebo differences strongly overlaps zero, since the posterior distribution for the 'true' difference does the same. The 95 per cent posterior prediction interval for the study-specific finasteride 5 mg-placebo difference does not exclude zero even though the posterior distribution for the 'true' difference does exclude zero. This reflects the fact that some of the studies included in Table I had negative finasteride 5 mg-placebo differences.

© 1998 John Wiley & Sons, Ltd.

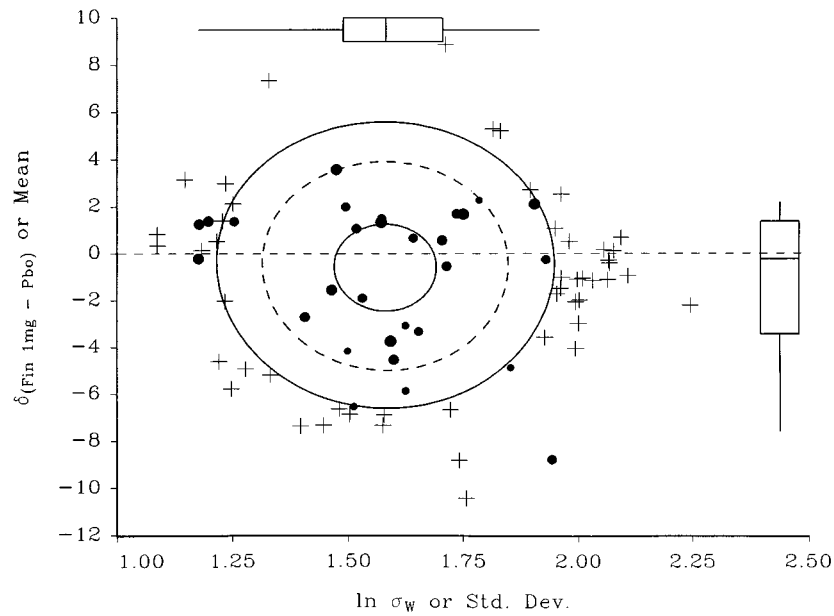

Figure 3. Bivariate box plot of joint posterior distribution of 5000 simulated mean differences (finasteride 1 mg - placebo) and within-group standard deviations, and observed outcomes from each centre

Figure 3 provides a graphical description of the Bayesian calculations analogous to the right-hand part of Figure 1 using a slightly modified version of a bivariate box plot.[31] The result is very similar to what the right-hand plot of Figure 1 shows; the observed difference distribution overlaps the distributions of the expected study-specific differences, as might be expected, so that the Bayesian approach also has provided a resonable description of the observed data.

DuMouchel describes some additional useful summaries of the Bayesian analysis.[9] Figure 4 displays the result of applying one of them to the data in Table I, to illustrate how a Bayesian analysis can shrink the individual study estimates towards a common mean. The degree of shrinkage depends on the heterogeneity of results among the studies. There would be little shrinkage if the heterogeneity were extensive. The most divergent findings tend to be those from the studies with the largest standard errors for the estimated difference, that is, the smallest studies, because the standard deviations vary little among the trials. The variability among the studies does not appear excessive, so we can conclude that true mean difference is zero or, at best, very slightly negative. The centre × treatment interaction detected by the fixed effect ANOVA therefore appears to represent variability among nearly zero outcomes rather than true interaction. This is the same conclusion that the other random effects analyses reach.

5.3. Distribution of other quantities

The flexibility of the Gibbs sampling methodology allows us to obtain the distributions of other interesting quantities. Thus, it is simple to obtain the posterior distribution of ξ, the expectation of $\ln \sigma_W^2$ for which the empirical Bayes approach calculates the maximum likelihood estimates. The

© 1998 John Wiley & Sons, Ltd.

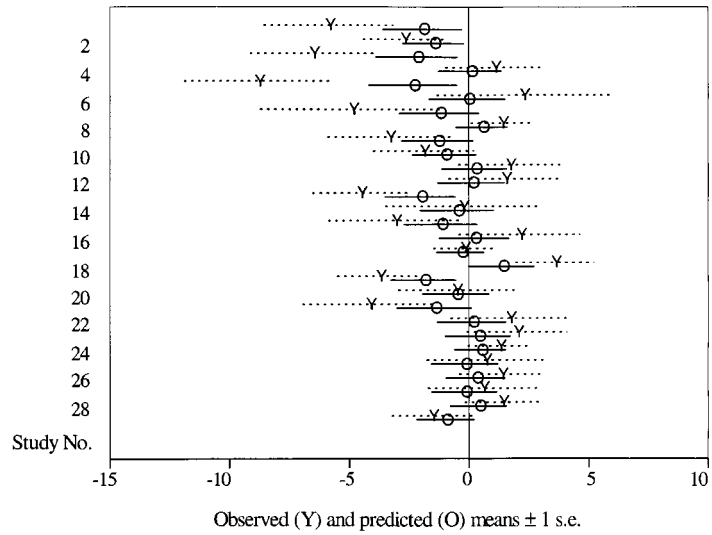

Figure 4. Observed and posterior predicted mean differences (finasteride 1 mg - placebo) $\pm$ 1 SE (observed differences) or 1 posterior SD (posterior predicted differences) for each study

expected value of ξ is 3·17, and the 95 per cent posterior probability interval is (3·04, 3·27). These values are very close to those from the empirical Bayes analysis (Table IV). Another interesting quantity is the posterior probability that the differences $\Delta_{\text{Fin1-Pbo}}$ and $\Delta_{\text{Fin5-Pbo}}$ both are negative, that is, that both dosages are effective, which we estimate by the fraction of simulations where both differences are negative. This probability turns out to be 0·83, suggesting that the evidence supports activity of both dosages. Whether the improvement realized with the 1mg dosage is clinically meaningful could be investigated easily by determining the posterior probability that the differences both are less than some minimally interesting value, for example, $-1{\cdot}0$. Yet another question worth considering is the likelihood that both dosages will turn out to be effective in any individual study, which we estimate by the fraction of simulations where the differences $\Delta_{\text{Fin1-Pbo}}$ and $\delta_{\text{Fin5-Pbo}}$ both are negative. This probability turns out to be about 0·59. It would be less if the differences both had to be less than a negative clinically meaningful value. Consequently, there is likely to be considerable variation among individual studies in whether both or only one dosage is shown effective.

5.4. Diagnostics

A variety of diagnostic tools may be applied to check the stability and precision of the computations. These include repeating the calculations with different starting estimates of the parameters, using different priors, checking smoothed moving averages of sampled values, and various analytic tools. The BUGS package provides a number of graphical and numerical diagnostics for assessing the stability of the computations. We summarize the performance of a few of these for the independence model; the results for the natural conjugate prior model are essentially the same. Taken together, the diagnostics suggest that stable estimates of the posterior distributions of the various parameters have been achieved, and, in particular, that the estimates

© 1998 John Wiley & Sons, Ltd.

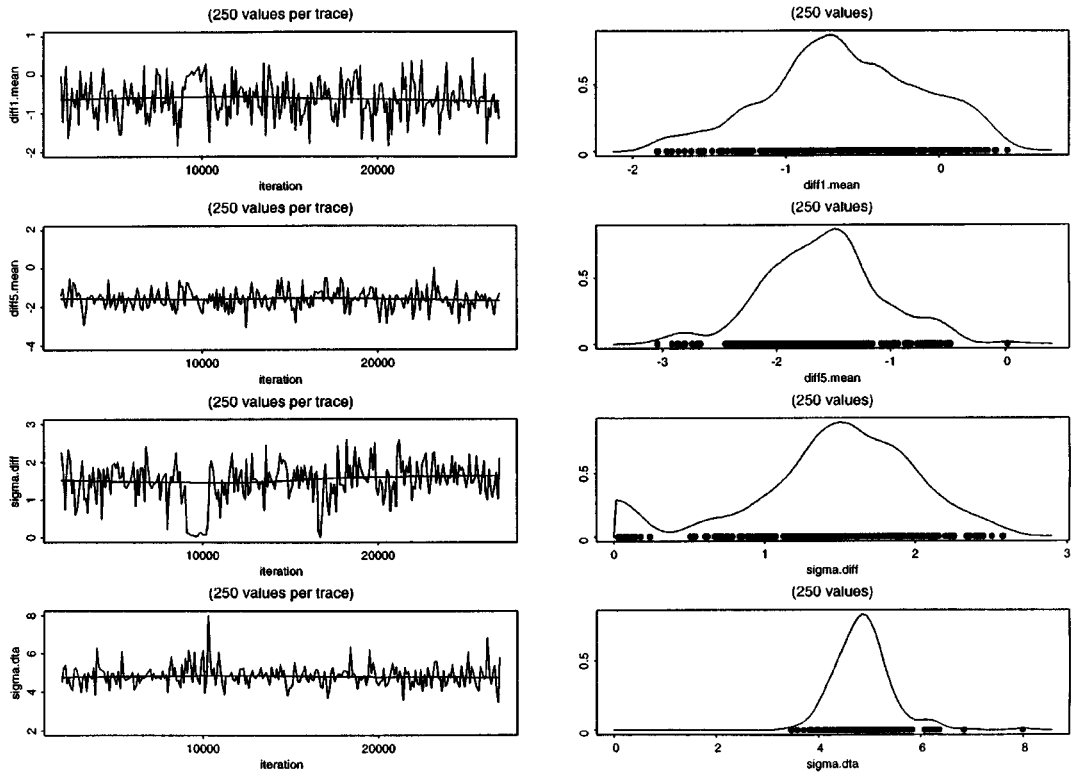

Figure 5. Traces of replicate values and kernel estimates of posterior densities of parameters in the model using the independence prior

of the means of the posterior distributions reasonably estimate the distributions' expected values. This conclusion is reinforced by the lack of dependence of the estimates on the starting values and the essentially identical results obtained from using two definitions for the prior distribution. Figure 5 displays the kernel estimates of the posterior distributions of $\Delta_{\text{Fin5-Pbo}}$, $\Delta_{\text{Fin1-Pbo}}$, σ_{W} [sigma.data], and σ_{B} [sigma.diff], along with traces of the values from the replicates after thinning and smoothed lines through the trace plots. The traces are relatively flat, which suggests that the simulated values are not drifting, that is, that stationarity may be appropriate. The kernel density estimates also seem reasonable, except possibly for the distribution of σ_{B}, which appears to be bimodal. The effect of this strange shape is to make precise estimation of the lower percentiles difficult, although the estimate of the mean is not greatly affected. This interpretation is supported by the results of the diagnostic procedures.

Gewecke describes a time-series-based diagnostic for evaluating the convergence of the means of posterior distributions that compares early iterated values with later ones.[32] The early and later values should be similarly distributed if early convergence of the iterates to a stationary sequence of samples from the posterior distribution has occurred. The values of Gewecke's convergence diagnostic, expressed as standard normal variate values, are less than 1·9 for all of the parameters ($\Delta_{\text{Fin5-Pbo}}$, $\Delta_{\text{Fin1-Pbo}}$, σ_{W}, σ_{B}), which implies that the process generating values from the posterior distributions has converged.

© 1998 John Wiley & Sons, Ltd.

Heidelberger and Welch describe another diagnostic for evaluating stationarity, including an evaluation of how many of the initial iterates to discard in order that the remaining iterates comprise a stationary sequence.[33] Their procedure also assesses the precision of the estimated posterior distribution mean in terms of the halfwidth of a suitably calculated 95 per cent confidence interval for the mean. If this halfwidth does not exceed a small fraction (for example 0·1) of the mean, then the process is judged to have estimated the mean sufficiently precisely. The sequences for all of the parameters satisfied the stationarity requirement, and only the sequence for σ_W needed to have initial iterates discarded (200/500 total iterates). The means of the posterior distributions were estimated with adequate precision except for the mean of the posterior distribution of $\delta_{\text{Fin1-Pbo}}$, for which the mean was $-0{\cdot}7$ and the confidence interval half width was 0·134. The ratio half width/mean equals 0·19, which exceeds 0·1, but is not so large that we would be inclined to conclude that the estimate was unacceptably imprecise.

Raftery and Lewis describe a diagnostic that also tests for stationarity of the sequence of iterates, but also checks how adequately the quantiles of the posterior distribution are estimated.[34] The BUGS defaults define 'adequate' as '95 per cent confident that the 2·5 per cent percentile has been estimated to a precision of $\pm$ 0·005'. The Raftery–Lewis procedure indicates how many burn-in iterations are required, what thinning interval would be appropriate to remove the effects in autocorrelation, and how many iterations are required. We applied the procedure to the run of 25,000 iterates without thinning, and found that for all parameters except for σ_B, a thinning interval of 2, a burn-in of 8 or fewer iterations (after thinning) and a total of less than 10,000 post-thinning iterations wo uld be sufficient to estimate the 2·5th percentile precisely enough. However, a thinning interval of 22, a burn-in of 440 post-thinning iterations, and more than 400,000 retained iterations would be required to estimate the 2·5th percentile of σ_B precisely enough. This is not surprising in view of the rather strange shape of the posterior distribution of σ_B shown in Figure 5.

6. DISCUSSION

The four strategies considered in this paper have provided increasingly insightful analyses at the cost of increasingly elaborate computations and output. However, all of the calculations can be carried out in an essentially production context, so that complex and costly programming is not needed each time a calculation has to be done. The essential differences are in the richness of the results, which require more thought and interpretation. Where they addressed the same questions and issues, the strategies led to the same conclusions. The ANOVA strategies were very simple to carry out, for the most part, but provided relatively limited information. The empirical and conventional Bayes strategies required modestly more effort to set up and interpret, but provided a lot of useful information about effect variability. This might be expected, since the ANOVA strategies are targeted primarily toward hypothesis testing, that is, whether an effect exists or not, while the emphasis with the Bayesian approaches is towards estimation and interval inference, that is, what can be said of the distributions of the effect sizes.

Whether centre and centre $\times$ treatment effects are regarded as fixed or random is central to the whole discussion. The calculations are aimed at inferences about treatment effects as manifested in multi-centre trials where the outcomes from the various centres may vary appreciably. The more or less conventional practice has been to apply factorial fixed effect ANOVA calculations with a term for treatment-by-centre interaction. This approach regards the centre and interaction effects as fixed, which simplifies the analysis but raises potential interpretation

© 1998 John Wiley & Sons, Ltd.

problems regardless of how the test for interaction turns out. Regarding the centre and interaction effects as random idealizes the mechanism by which these effects are manifested, but also can provide a more realistic assessment of the uncertainty associated with the inferences about treatment effects. If, in the context of a fixed effects model, the interaction effect turns out to be significant in the usual sense, then the treatment effect differences might not be interpretable meaningfully. If the effect is not significant, then a test for a treatment effect might be carried out using only the within-group mean square; this is correct in a narrow technical sense, but it does underestimate the variability that really affects the responses to treatment. Random effect models lose the ability to identify treatment-by-centre interaction effects explicitly, at least formally, but it is not clear that this interaction has much meaning in the usual context of multi-centre trials as opposed to the factorial design circumstances for which it is intended. Regardless of the model, centres whose outcomes differ markedly from those of the bulk of the centres, that is, apparent outliers, need to be followed up to determine why, since the effects of treatments applied in a consistent manner to similar kinds of patients should be manifested consistently regardless of the centre in which the observations are made.

Methods that both incorporate random variation among centres explicitly and provide a way to identify possible outliers therefore provide some advantages that may offset the possibly more complex computations they require. The empirical and conventional Bayes methods described here comprise examples of such methods. The empirical Bayes approach is particularly appealing in this regard, since it provides a way to test the goodness-of-fit of the presumed model and also a way to identify potential outliers. The conventional Bayes approach based on Gibbs sampling provides a very simple way to carry out inferences on functions of the data and the parameters that would be very difficult to do in any other way. The effect of changing distributional assumptions also can be studied easily. For example, instead of assuming that the parameters are drawn from normal distributions, we might evaluate the consequences of assuming that the parameters are drawn from distributions with heavier tails such as a central t with 3 degrees of freedom. The Bayesian methods have the additional advantage of being likelihood based and therefore more robust to the effect of outliers than methods such as ANOVA based on squared errors that can be strongly influenced by outliers.

The level of sophistication and insight provided by empirical and conventional Bayesian methods probably would not justify the effort required to implement them in all circumstances, although we emphasize that this effort is not a great deal more than that required by an ANOVA; the extra effort comes in interpreting the extra information these procedures supply. However, it would seem very worthwhile to employ them for key outcome variables where insight into the outcome of the trial is particularly important.

The emphasis on estimation and interval inference that characterizes the empirical and conventional Bayesian approaches can be particularly valuable in designing subsequent trials. Multi-centre trials of a compound often provide a foundation for future trials in that compound or in future compounds. Not all of these trials will be aimed at detecting differences from placebo or a relatively inactive control. Whether a subsequent trial is aimed at detecting a difference between a test agent and a control, or is aimed at establishing the equivalence of two agents in some clinically meaningful sense, a description of the anticipated variability and parameter values such as the initial uncertainty profile[35] is essential in determining the sample size. The empirical and conventional Bayes approaches provide this description explicitly.

© 1998 John Wiley & Sons, Ltd.

APPENDIX: ANNOTATED SETUP FOR BUGS RUNS

```
{
  tau.diff ~ dgamma (1·0E − 3, 1·0E − 3);                Prior distributions
  diff1.mean ~ dnorm (0·0, 1·0E − 5);
  diff5.mean ~ dnorm (0·0, 1·0E − 5);
  for (j in 1: nomega) {omcat [j] ← 1/nomega};           Use categorization because
  for (j in 1: nomega) {zetcat [j] ← 1/nomega};          log-concavity fails - uniform
  om ~ dcat (omcat [ ]);                                 prior distribution over categories
  zet ~ dcat (zetcat [ ]);
  zeta ← 9 + 1.33*zet;                                   Rescale to reflect anticipated
  omega2 ← 5*om;                                         values
  omz ← omega2*zeta;

  for (i in 1: nstdy) {                                  Loop calculations over centres
    tau.dta [i] ~ dgamma (omega2, omz);                  Random variance each centre
    z [i] ← 0.5*tau.dta [i];

    df.fin1 [i] ← n.fin1 [i] − 1;                        Observed d.f. and MSE in the
    ms2.fin1 [i] ← df.fin1 [i]*pow (std.fin1 [i],2);     finasteride 1mg group

        … same steps for Fin5 and placebo groups …

    df.diff [i] ← df.fin1 [i] + df.fin5 [i] + df.pbo [i];     d.f. for pooled MSE
    diff.prm [i] ← 0.5* df.diff [i];                          Obs'd pooled MSE
    ms2.diff [i] ~ dgamma (diff.prm [i],z [i]);               Dist'n of pooled MSE

    ms2.diff [i] ← ms2.fin1 [i] + ms2.fin5 [i] + ms2.pbo [i];     Pooled within SS

    n.diff1 [i] ← 1·0/ (1·0/n.fin1 [i] + 1.0/n.pbo [i]);      Effective sample size for
    prec.diff1 [i] ← n.diff1 [i]*z [i];                       fin.1mg - pbo difference
    mean.diff1 [i] ~ dnorm (diff1.mean, tau.diff);            Dist'n of centre mean of
                                                               fin.1mg - pbo difference
    obs.mean.diff1 [i] ← mean.fin1 [i] − mean.pbo [i];          Observed difference
    obs.mean.diff1 [i] ~ dnorm (mean.diff1 [i], prec.diff1 [i]);   Dist'n of observed
                                                                   differences
        … same steps for Fin5 group …
  }                                                            End loop over centres
  tau.study.dta ~ dgamma(omega2, omz);
  sigma.dta ← 1/sqrt(tau.study.dta);
  sigma.diff ← 1/sqrt(tau.diff);
  omega ← 2*omega2;

  diff1.study.mean ~ dnorm (diff1.mean, tau.diff);       Sample from predictive
  diff5.study.mean ~ dnorm (diff5.mean, tau.diff);       dist'ns for centre-specific
}                                                        mean differences
update(2000)                          Burn-in runs
update(25000)                         Runs with results retained
                                                                          Initial
inits()    list(diff1.mean = 0, diff5.mean = 0, om = 1, zet = 10, tau.diff = 1)   values
```

© 1998 John Wiley & Sons, Ltd.

REFERENCES

1. Fleiss, J. L. 'Analysis of data from multiclinic trials', *Controlled Clinical Trials*, **7**, 267–275 (1986).
2. Ciminera, J. L., Heyse, J. F., Nguyen, H. H. and Tukey, J. W. 'Tests for qualitative treatment-by-centre interaction using a "pushback" procedure', *Statistics in Medicine*, **12**, 1033–1045 (1993).
3. Ciminera, J. L., Heyse, J. F., Nguyen, H. H. and Tukey, J. W. 'Evaluation of multicentre clinical trial data using adaptations of the Mosteller-Tukey procedure', *Statistics in Medicine*, **12**, 1047–1061 (1993).
4. Smith, T. C., Spiegelhalter, D. J. and Thomas, A. 'Bayesian approaches to random-effects meta-analyses: A comparative study', *Statistics in Medicine*, **14**, 2685–2699 (1995).
5. Whitehead, A. and Whitehead, J. 'A general parametric approach to the meta-analysis of randomized clinical trials', *Statistics in Medicine*, **10**, 1665–1677 (1991).
6. van Houwelingen, H. C., Zwinderman, K. H. and Stijnen, T. 'A bivariate approach to meta-analysis', *Statistics in Medicine*, **12**, 2273–2284 (1993).
7. Hardy, R. J. and Thompson, S. G. 'A likelihood approach to meta-analysis with random effects', *Statistics in Medicine*, **15**, 619–629 (1996).
8. Goldstein, H. *Multilevel Statistical Models*, Halstead Press, New York, 1995.
9. DuMouchel, W. 'Predictive cross-validation of Bayesian meta-analyses', *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics 5*, Oxford University Press, Oxford, 1996, pp. 107–127.
10. Gormley, G. J., Stoner, E., Bruskewitz, R. C., Imperato-McGinley, J., Walsh, P. C., McConnell, J. D., Andriole, G. L., Geller, J., Bracker, B. R., Tenover, J. S., Vaughan, E. D., Pappas, F., Taylor, A., Binkowitz, B. and Ng, J. 'The effect of finasteride in men with benign prostatic hyperplasia', *New England Journal of Medicine*, **327**, 1185–1191 (1992).
11. *SAS 6.10.* SAS Institute, Inc. Cary, NC, 1995.
12. Gail, M. and Simon, R. 'Testing for qualitative interactions between treatment effects and patient subsets', *Biometrics*, **41**, 361–372 (1985).
13. Harville, D. A. 'Maximum likelihood approaches to variance component estimation and to related problems', *Journal of the American Statistical Association*, **72**, 320–340 (1977).
14. Searle, S. R., Casella, G. and McCulloch, C. E. *Variance Components*, Wiley, New York, 1992.
15. *S-Plus 3.3.* Statistical Sciences, Inc, Seattle, WA, 1995.
16. Maritz, J. S. and Lwin, T. *Empirical Bayes Methods*, Chapman and Hall, New York, 1989.
17. Morris, C. N. 'Parametric empirical Bayes inference', *Journal of American Statistical Association*, **78**, 47–65 (1983).
18. Kass, R. E. and Steffey, D. 'Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)', *Journal of American Statistical Association*, **84**, 717–726 (1989).
19. O'Hagan, A. *Kendall's Advanced Theory of Statistics, Vol. 2B. Bayesian Inference*, Halstead Press, New York, 1994.
20. Bates, D. M. and Watts, D. G. *Nonlinear Regression Analysis and Its Applications*, Wiley, New York, 1998.
21. Khamis, H. J. 'The d-corrected Kolmogorov-Smirnov test for goodness-of-fit', *Journal of Statistical Planning and Inference*, **24**, 317–335 (1990).
22. Leurgans, S. 'On distribution-free tests for bivariate observations', *Biometrika*, **70**, 727–728 (1983).
23. *GAUSS 3.2.* Aptech Systems, Inc. Maple Valley, WA, 1995.
24. Carlin, B. P. and Gelfand, A. E. 'Approaches for Empirical Bayes confidence intervals', *Journal of American Statistical Association*, **85**, 105–114 (1990).
25. Kass, R. E. and Raftery, A. E. 'Bayes factors', *Journal of American Statistical Association*, **90**, 773–795 (1995).
26. Raftery, A. E. 'Approximate Bayes factors and accounting for model uncertainty in generalized linear models', Technical Report 255, University of Washington, Department of Statistics, 1993.
27. Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. 'Modelling complexity: Applications of Gibbs sampling in medicine (with discussion)', *Journal of the Royal Statistical Society, Series B*, **55**, 39–102 (1993).
28. Besag, J., Green, P., Higdon, D. and Mengersen, K. 'Bayesian computation and stochastic systems (with discussion)', *Statistical Science*, **10**, 3–66 (1995).

© 1998 John Wiley & Sons, Ltd.

29. *BUGS 0.3*, MRC Biostatistics Unit, Cambridge University, Cambridge, 1995.
30. Fisher, L. D. 'Comments on Bayesian and frequentist analysis and interpretation of clinical trials', *Controlled Clinical Trials*, **17**, 423–434 (1996).
31. Goldberg, K. M. and Iglewicz, B. 'Bivariate extensions of the boxplot', *Technometrics*, **34**, 307–320 (1992).
32. Gewecke, J. 'Evaluating the accuracy of sampling-based approaches to calculating posterior moments', *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds.) *Bayesian Statistics 4*, Clarendon Press, Oxford, U.K., 1992.
33. Heidelberger, P. and Welch, P. 'Simulation run length control in the presence of an initial transient', *Operations Research*, **31**, 1109–1144 (1983).
34. Raftery, A. E. and Lewis, S. 'How many iterations in the Gibbs sampler?', *in* Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics 4*, Clarendon Press, Oxford, U.K., 1992.
35. Gould, A. L. 'Sample sizes for event rate equivalence trials using prior information', *Statistics in Medicine*, **12**, 2009–2023 (1993).

© 1998 John Wiley & Sons, Ltd.