BIOMETRICS 53, 456–464
June 1997

Bayesian Design and Analysis of Two × Two Factorial Clinical Trials

Richard Simon

Biometric Research Branch, National Cancer Institute,
Bethesda, Maryland 20852, U.S.A.

and

Laurence S. Freedman

Biometry Branch, National Cancer Institute,
Bethesda, Maryland 20852, U.S.A.

SUMMARY

The 2 × 2 factorial design has been advocated for improving the efficiency of clinical trials. Most such trials are designed on the assumption that there is no interaction between the levels of the factors and outcome. This assumption is often problematic, however, because interactions are usually possible in clinical trials and the sample sizes often used provide little power in testing for interactions. We consider the use of Bayesian methods for the design and analysis of 2×2 factorial clinical trials. This approach avoids the need to dichotomize one's assumptions that interactions either do or do not exist and provides a flexible approach to the design and analysis of such clinical trials. Exact results are developed for balanced factorial designs with normal response. Approximations are then presented for factorial designs based on the logistic model for binary response or the proportional hazards model for time-to-event data. The resulting approximate posterior distributions are normal and hence no extensive computations are required. Suggestions for specification of prior distributions are presented.

1. Introduction

Factorial designs have long been recognized in agricultural and industrial experimentation as an important tool for studying the joint effect of several factors and for selecting factors for further experimentation (Fisher, 1935). More recently, factorial designs have been recommended as an approach to increasing the efficiency of clinical trials (Buring and Hennekens, 1990; Byar and Piantadosi, 1985; Peto et al., 1976; Stampfer et al., 1985). Phase III clinical trials often represent a setting for confirmatory testing of treatments rather than for exploratory screening, however. The results of phase III trials are often used as a basis for drug approval or for national health care recommendations. There has been controversy over the use of factorial designs in such settings. Although factorial designs are clearly appropriate for studying the joint effects of several factors, there is controversy over the appropriate size of such studies. Proponents have suggested that the sample size for a factorial design be based on an assumption that no interaction exists between the two factors. In treatment studies, where the factors are treatments with two levels (given or not given), it is often important to know whether the treatments should be given together. Both positive and negative interactions are possible. If the factorial design is sized based on the assumption that no interaction exists, it will not be large enough to provide a statistically powerful test of the hypothesis of no interaction (Peterson and George, 1993). Consequently, factorial designs are often avoided unless one can assume beforehand that there is no interaction among the factors (Brittain and Wittes, 1989). In many cases, it seems unnatural to be forced to assume that interactions do

Key words: Bayesian analysis; Clinical trials; Factorial design.

or do not exist. We have therefore investigated a Bayesian formulation of the 2×2 factorial design that does not force this dichotomous decision in planning the trial or analyzing the results.

2. The Model

Consider the 2×2 factorial design and let x_1, x_2, y denote the level of factor one, factor two, and the value of the response, respectively. The factors will each be coded -1 or 1 for the low level (treatment not given) and high level (treatment given). The model that we will use is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$, where the error terms ϵ are independently normal with mean zero and variance σ^2. With this parameterization, the main effects of treatment 1 and 2 are $2\beta_1$ and $2\beta_2$, respectively, where main effect means the expected response at the high level of a factor minus the expected response at the low level of the factor averaged over the levels of the other factor. The difference between the effect of treatment one at the high level of treatment 2 and the effect of treatment one at the low level of treatment 2 equals $4\beta_3$. We assume that σ^2 is known and that there are n observations in each of the four cells determined by the levels of the two treatments.

The maximum likelihood estimator (mle) $\underline{\hat{\beta}}$ of $\underline{\beta}$ is

$$\begin{aligned}
\hat{\beta}_0 &= (y_{\cdot} + y_1 + y_2 + y_{12})/4 \\
\hat{\beta}_1 &= (y_{12} + y_1 - y_2 - y_{\cdot})/4 \\
\hat{\beta}_2 &= (y_{12} + y_2 - y_1 - y_{\cdot})/4 \\
\hat{\beta}_3 &= (y_{12} + y_{\cdot} - y_1 - y_2)/4,
\end{aligned}$$

where y_1 denotes the mean of the observations with treatment one at its upper level and treatment 2 at its lower level, y_2 denotes the mean of the observations with treatment 2 at its upper level and treatment one at its lower level, y_{12} denotes the mean of the observations with both treatments at their upper levels, and $y_{\cdot}$ denotes the mean of the observations with both treatments at their lower levels. It is easily shown that $\underline{\hat{\beta}}$ is normal with mean $\underline{\beta}$ and variance $(\sigma^2/4n)I$.

3. Posterior Distributions

We will assume that the prior distribution for $\underline{\beta}$ is $\mathrm{N}(\underline{\mu}, \Sigma)$. Lindley and Smith (1972) showed that with such a prior distribution, if $\underline{\hat{\beta}} \mid \underline{\beta} \sim \mathrm{N}(\underline{\beta}, C)$, then the posterior distribution is given by $\underline{\beta} \mid \underline{\hat{\beta}} \sim \mathrm{N}(B\underline{b}, B)$, where $B^{-1} = C^{-1} + \Sigma^{-1}$ and $\underline{b} = C^{-1}\underline{\hat{\beta}} + \Sigma^{-1}\underline{\mu}$. We note that $\underline{\beta} \mid \underline{\hat{\beta}}$ is the posterior distribution for the parameters since the mle $\underline{\hat{\beta}}$ is a sufficient statistic. As noted above for this model, $\mathbf{C} = (\sigma^2/4n)\mathbf{I}$. We will take $\Sigma = \mathrm{diag}(\lambda^2, \tau_1^2, \tau_2^2, \nu^2)$. This assumes that our prior distributions for the four parameters are independent. Dependent priors would permit one to specify that large main effects for one factor would tend to be associated with large main effects of the other factor or that large interactions tend to be associated with large main effects. Here we will focus on the simpler case of independent priors. λ^2 is the variance of the prior for the intercept β_0, τ_1^2 and τ_2^2 are the variances of the priors for the main effects of the two treatments, and ν^2 is the variance of the prior for the interaction. Under these conditions, the posterior distribution of $\underline{\beta}$ is normal with mean

$$\begin{pmatrix}
(4n/\sigma^2 + 1/\lambda^2)^{-1} & ((4n/\sigma^2)\hat{\beta}_0 + (1/\lambda^2)\mu_0) \\
(4n/\sigma^2 + 1/\tau_1^2)^{-1} & ((4n/\sigma^2)\hat{\beta}_1 + (1/\tau_1^2)\mu_1) \\
(4n/\sigma^2 + 1/\tau_2^2)^{-1} & ((4n/\sigma^2)\hat{\beta}_2 + (1/\tau_2^2)\mu_2) \\
(4n/\sigma^2 + 1/\nu^2)^{-1} & ((4n/\sigma^2)\hat{\beta}_3 + (1/\nu^2)\mu_3)
\end{pmatrix}, \tag{1}$$

which is a weighted average of the mle and the prior mean with weights being the reciprocals of the variances of the mle and the prior. The posterior variance of $\underline{\beta}$ is

$$\begin{pmatrix}
(4n/\sigma^2 + 1/\lambda^2)^{-1} & 0 & 0 & 0 \\
0 & (4n/\sigma^2 + 1/\tau_1^2)^{-1} & 0 & 0 \\
0 & 0 & (4n/\sigma^2 + 1/\tau_2^2)^{-1} & 0 \\
0 & 0 & 0 & (4n/\sigma^2 + 1/\nu^2)^{-1}
\end{pmatrix}. \tag{2}$$

If noninformative priors are used for all of the components of $\underline{\beta}$, that is $\lambda^2, \tau_1^2, \tau_2^2, \nu^2 \to \infty$, then the posterior distribution has mean $\underline{\hat{\beta}}$ and variance $(\sigma^2/4n)\mathbf{I}$, the same variance as the mle. If we use noninformative priors for the intercept and main effects but assume that there is no interaction

(i.e., $\mu_3 = \nu^2 = 0$), then the posterior distribution of $\underline{\beta}$ has mean $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, 0)$ and variance $(\sigma^2/4n)\text{diag}(1,1,1,0)$. Both of these special cases represent extreme assumptions, however. The first corresponds to the assumption that interactions of all sizes are equally plausible *a priori*. The second corresponds to the assumption that interactions of any size other than zero are completely implausible. This second case is often used in designing factorial trials.

4. Analysis of Clinical Trial Results

The analysis of a 2×2 factorial clinical trial can be performed using the above results. Because the posterior distribution of $\underline{\beta}$ is normal and the posterior covariance matrix is diagonal, the analysis is particularly simple. The posterior means and variances for the effects of usual interest are shown in Table 1. It should be remembered that the main effect of treatment j ($j = 1$ or 2) is $2\beta_j$ and that this represents the effect of the treatment averaged over the levels of the other treatment. The difference between the effect of treatment 1 when given with treatment 2 and the effect of treatment 1 when given alone is $4\beta_3$.

In addition to examining the posterior distribution of the main effects of each treatment and of the interaction parameter, the effects of each treatment within fixed levels of the other treatment will also be of interest. The effect of treatment 1 when given alone is $2\beta_1 - 2\beta_3$, whereas the effect of treatment 1 when given with treatment 2 is $2\beta_1 + 2\beta_3$. Similarly, the effects of treatment 2 when given alone or with treatment 1 are $2\beta_2 - 2\beta_3$ and $2\beta_2 + 2\beta_3$, respectively. The posterior means of these within-level estimates of treatment effect can be computed by adding or subtracting the appropriate elements of the posterior mean vector given above in (1) and multiplying by two.

Although the posterior distributions of the main effects and the interaction are independent of each other, this is obviously not true for the posterior distributions of the effect of one treatment within each of the two levels of the other treatment. It is easily shown, however, that the covariance between the effects of treatment j at the two levels of the other treatment is $4\,\text{var}(\beta_j) - 4\,\text{var}(\beta_3)$, where the variance refers to the posterior distribution. Hence, one can easily make joint probability statements about the effects of a treatment within the two levels of the other treatment.

Inferences can be easily derived using the results shown in Table 1. For example, a 95% highest posterior density interval for the main effect of treatment 1 ($2\beta_1$) is

$$2\frac{4n\hat{\beta}_1/\sigma^2 + \mu_1/\tau_1^2}{4n/\sigma^2 + 1/\tau_1^2} \pm 3.92(4n/\sigma^2 + 1/\tau_1^2)^{-0.5}.$$

The posterior probability that the main effect of treatment 1 is negative is

$$\Phi\left(-\frac{4n\hat{\beta}_1/\sigma^2 + \mu_1/\tau_1^2}{(4n/\sigma^2 + 1/\tau_1^2)^{0.5}}\right),$$

Table 1

Posterior means and variances for treatment effects of usual interest

	Parameter	Posterior mean	Posterior variance
Main effect of treatment 1	$2\beta_1$	$2\dfrac{k\hat{\beta}_1 + \mu_1/\tau_1^2}{k + 1/\tau_1^2}$	$\dfrac{4}{k + 1/\tau_1^2}$
Main effect of treatment 2	$2\beta_2$	$2\dfrac{k\hat{\beta}_2 + \mu_2/\tau_2^2}{k + 1/\tau_2^2}$	$\dfrac{4}{k + 1/\tau_2^2}$
Interaction	$4\beta_3$	$4\dfrac{k\hat{\beta}_3 + \mu_3/\nu^2}{k + 1/\nu^2}$	$\dfrac{16}{k + 1/\nu^2}$
Effect of treatment 1 without treatment 2	$2\beta_1 - 2\beta_3$	$2\dfrac{k\hat{\beta}_1 + \mu_1/\tau_1^2}{k + 1/\tau_1^2} - 2\dfrac{k\hat{\beta}_3 + \mu_3/\nu^2}{k + 1/\nu^2}$	$\dfrac{4}{k + 1/\tau_1^2} + \dfrac{4}{k + 1/\nu^2}$
Effect of treatment 2 without treatment 1	$2\beta_2 - 2\beta_3$	$2\dfrac{k\hat{\beta}_2 + \mu_2/\tau_2^2}{k + 1/\tau_2^2} - 2\dfrac{k\hat{\beta}_3 + \mu_3/\nu^2}{k + 1/\nu^2}$	$\dfrac{4}{k + 1/\tau_2^2} + \dfrac{4}{k + 1/\nu^2}$

$k = 4n/\sigma^2$.

where Φ denotes the standard normal distribution function. Interval estimation and hypothesis testing for other functions of parameters can be similarly computed based on the normal distribution and the posterior means and variances.

5. Design Considerations

Design considerations include specifying the prior distributions and planning the number of patients to be accrued. The usual frequentist approach is analogous to the use of flat priors on the intercept and main effects ($\lambda^2, \tau_1^2, \tau_2^2 \to \infty$) and the concentration of the prior for interaction at zero ($\mu_3 = 0, \nu^2 = 0$). Using flat priors for the intercept and main effects is often viewed as making the analysis objective and independent of prior beliefs, but such priors can also be viewed as representations of extreme beliefs that all values are equally likely. Often there will be information from other trials for the treatments under consideration or for similar treatments that indicate that large positive or negative effects are less likely than smaller effects. Spiegelhalter, Freedman, and Parmar (1994) have advocated the use of "skeptical priors" for treatment effects to represent the views of consumers of research who may be less optimistic about the treatments than those investigators conducting the clinical trial. In this context, the priors for the main effects might be taken as having mean zero or mean halfway between zero and the smallest size of effect considered medically worthwhile in light of the costs, toxicity, and inconvenience of the treatments. The variances of the skeptical priors for the main effects can be calibrated to provide a specified probability that the treatment effect is greater than the smallest size considered medically worthwhile. The size of effect considered medically worthwhile is elicited from physicians. In another approach, the entire prior for the main effect can be elicited from physicians not associated with the clinical trial.

If there is no prior reason to expect an interaction in one direction more than the other, then the prior mean for the interaction term may be taken as zero ($\mu_3 = 0$). In some cases, negative interactions ($\beta_3 < 0$) may be considered more likely than positive interactions. If two drugs function via the same biological mechanism, then there may be no extra benefit for the combination. Negative interactions also result if the administration of one treatment interferes with the administration of the other treatment. Usually, however, this should be determined in a small pilot trial and a factorial design avoided in such circumstances. In many cases, little is known about the biological mechanism and both negative and positive interactions are possible. In such cases, a prior mean $\mu_3 = 0$ is reasonable.

The prior variance ν^2 is used to provide a subjective specification of the probability that the effects of one treatment depend on the level of the other treatment. This can also be elicited. For example, one might elicit the probability that the effect of treatment 1 in the presence of treatment 2 is at least Δ, given that there is no effect of treatment 1 when given alone. This probability of a qualitative interaction can be written

$$\Pr(2\beta_1 + 2\beta_3 \geq \Delta \mid 2\beta_1 - 2\beta_3 = 0) = \frac{\displaystyle\int_{\Delta/4}^{\infty} \phi\left(\frac{z-\mu_1}{\tau_1}\right)\phi\left(\frac{z-\mu_3}{\nu}\right)dz}{\displaystyle\int_{-\infty}^{\infty} \phi\left(\frac{z-\mu_1}{\tau_1}\right)\phi\left(\frac{z-\mu_3}{\nu}\right)dz}, \tag{3}$$

where ϕ is the probability density function for the standard normal distribution. With μ_1, μ_3, Δ, and τ_1 specified, this quantity is a function of ν. Hence, the prior variance of the interaction term can be determined to satisfy the value of the conditional probability elicited. In this process, Δ may be taken as the minimal medically worthwhile difference.

If the primary purpose of the trial is estimation of the main effects of the treatments, then the posterior variances of β_1 and β_2 are the quantities of interest for sizing the study. It is of interest to note in Table 1 that the posterior distributions of the main effects do not depend on the prior distribution for the interaction nor on the mle of the interaction. This indicates that the main effect can be estimated without assuming that there is no interaction and that the mle for the interaction does not influence our inference about the main effects. If we want to estimate $2\beta_1$ with a 95% posterior density interval of width Δ, then we should select n so that 1.96 times the posterior standard deviation equals Δ. That is,

$$n = \left(\frac{15.36}{\Delta^2} - \frac{1}{4\tau_1^2}\right)\sigma^2. \tag{4}$$

In expression (4), the constant 15.36 represents $4z^2_{(1-\gamma)/2}$, where γ is the probability level. For example, for $\gamma = .95, z_{.025} = -1.96$. Although the posterior distributions of the main effects are

independent of the prior for the interaction and independent of the mle of the interaction, the main effects will have less medical meaning when there is a large interaction.

If one wishes to ensure that the sample size is adequate to provide precise estimates of the effect of a factor within the levels of the other factor, then the posterior variances of $2\beta_1 \pm 2\beta_3$ and $2\beta_2 \pm 2\beta_3$ are the relevant quantities. As shown in Table 1, these posterior variances do involve the prior for the interaction. One can assure a sample size sufficient to estimate within-level treatment effects by 95% posterior density intervals of width Δ by selecting n so that 2 times 1.96 times the posterior standard deviation equals Δ. Solving the resulting quadratic equation gives

$$n = 2 * 1.96^2 * (\sigma^2/\Delta^2)\left[2 - k(1/\tau_1^2 + 1/\nu^2) + \sqrt{k^2(1/\tau_1^2 + 1/\nu^2)\left(1 - \frac{4}{\tau_1^2 + \nu^2}\right) + 4}\right], \quad (5)$$

where $k = (\Delta/(4 * 1.96))^2$.

For τ_1^2, $\nu^2 \to \infty$, the value of n obtained from (5) is twice that which would be obtained from (4) if we set $\tau_1^2 \to \infty$. This is because setting $\nu^2 = \infty$ and $\tau_1^2 = \infty$ corresponds to a frequentist approach, and this requires twice the sample size for estimating within-level treatment effects as for estimating main effects with the same precision. This is not the case for the general Bayesian analysis, however. The required sample size is a value intermediate between those obtained under the assumption that $\nu^2 = 0$ and under the assumption that $\nu^2 \to \infty$.

6. Extensions

The development above applies to balanced factorial designs with homoscedastic normal responses. In this section, we shall describe several extensions of these results.

First, suppose that the normal response structure is retained but the sample sizes are unequal. In this case, the expressions for the maximum likelihood estimators shown in Section 2 remain unchanged, but the covariance matrix becomes $(\sigma^2/4n_h)\mathbf{I}$, where n_h denotes the harmonic mean of the four sample sizes. All other results are as before with n_h replacing n.

6.1 *Logistic Models*

Binary response factorial experiments can be analyzed using the logistic model

$$\log\left(\frac{p(\underline{x})}{1 - p(\underline{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

where $p(\underline{x})$ denotes the probability of response associated with factors $\underline{x}$. The maximum likelihood estimator $\underline{\hat{\beta}}$ can be found in the usual manner and is asymptotically normal with mean $\underline{\beta}$ and covariance matrix $\mathbf{C}$ estimated as the inverse of the sample information matrix. In general, $\mathbf{C}$ will not be diagonal. Nevertheless, the posterior distribution of β can be approximated as multivariate normal with mean $B\underline{b}$ and variance B as defined previously. The approximation is based on the asymptotic sufficiency of the mle and its asymptotic normality.

Another approximate analysis of the binary response design is obtained using the empirical logistic transform. Results for each of the four treatment groups is summarized by the log of the odds of response

$$y = \log\left(\frac{\hat{p}(\underline{x})}{1 - \hat{p}(\underline{x})}\right),$$

where $\hat{p}(\underline{x})$ is the observed proportion of successes in the cell characterized by $\underline{x}$ and its approximate variance is

$$\mathrm{var}(y) \cong \frac{1}{n(\underline{x})\hat{p}(\underline{x})(1 - \hat{p}(\underline{x}))}, \quad (6)$$

where $n(\underline{x})$ is the sample size in the cell (Cox and Snell, 1989). The harmonic mean of these variances is used instead of σ^2/n in the formulas provided previously for analysis of the trial. For purposes of planning, a common estimate of $p(\underline{x})(1 - p(\underline{x}))$ is used because this will vary only slightly for typical treatment effects if the success probabilities are not close to zero or one.

6.2 *Proportional Hazards Models*

Proportional hazards models for survival data can be handled in a manner analogous to that described for logistic models. We will write the hazard function as $\lambda(t, \underline{x}) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$ using the parameterization previously described. The model is fit in the usual way,

providing the maximum partial likelihood estimator $\hat{\underline{\beta}}$ and the estimate $\mathbf{C}$ of the asymptotic covariance matrix. Analysis is based on the multivariate normal posterior distribution with mean $B\underline{b}$ and covariance matrix $\mathbf{B}$. Here the dimension of the mle is three rather than four, and the covariance matrix $\mathbf{C}$ may not be diagonal, but this causes no problems for analysis.

For planning proportional hazards factorial designs, we note that at the null hypothesis $\beta = (0, 0, 0)$, the sample information matrix has (k, l) element

$$\sum_{i \in U} \left(\frac{1}{n_i} \sum_{j \in R_i} (x_{jk} - \bar{x}_{(i)k})(x_{jl} - \bar{x}_{(i)l}) \right),$$

where U is the set of indices of cases with uncensored survivals, R_i is the risk set of the ith observation, n_i is the number of elements in R_i, x_{uv} is the value of the vth component of the vector x_u for case u, and

$$\bar{x}_{(i)v} = \frac{1}{n_i} \sum_{j \in R_i} x_{jv}.$$

Consequently, as noted by Miller (1981), the sample information matrix is simply a sum of the covariate covariance matrices for the risk sets of the uncensored observations. For a balanced factorial design under the null hypothesis, the sample information matrix will be close to diagonal. A similar analysis for the full exponential likelihood would show that the sample information matrix for a balanced factorial design under the null hypothesis is exactly diagonal. For the proportional hazards model with large samples, the diagonal elements approximately equal n_u, the number of uncensored observations. Hence, the covariance matrix of $\hat{\underline{\beta}}$ may be approximated by $(1/n_u)\mathbf{I}$ for purposes of study planning.

If the prior distribution for $\underline{\beta}$ is multivariate normal with mean (μ_1, μ_2, μ_3) and diagonal covariance matrix with diagonal elements $(\tau_1^2, \tau_2^2, \nu^2)$, then the posterior distribution of $\underline{\beta}$ can be approximated for planning purposes by a multivariate normal distribution with mean

$$\begin{bmatrix} \dfrac{n_u\hat{\beta}_1 + \mu_1/\tau_1^2}{n_u + 1/\tau_1^2} \\ \dfrac{n_u\hat{\beta}_2 + \mu_2/\tau_2^2}{n_u + 1/\tau_2^2} \\ \dfrac{n_u\hat{\beta}_3 + \mu_3/\nu^2}{n_u + 1/\nu^2} \end{bmatrix}$$

and diagonal covariance matrix. The posterior variances of β_1, β_2, and β_3 are $(n_u + 1/\tau_1^2)^{-1}$, $(n_u + 1/\tau_2^2)^{-1}$, and $(n_u + 1/\nu^2)^{-1}$, respectively. Hence, for planning purposes with the proportional hazards model, the results are similar to the normal linear model using $\sigma^2 = 4, n = n_u$, and no intercept term.

7. Example

Eisenhauer et al. (1994) reported a clinical trial of two dose levels of taxol and two schedules of administration for patients with advanced ovarian carcinoma whose disease had already progressed in the presence of firstline chemotherapy. The trial was planned as a 2×2 factorial assuming that there would be no interaction. The trial organizers intended to randomize a total of 300 patients, but indicated that they continued accrual to 407 patients because of an initial imbalance in numbers between the 3- and 24-hour groups. Of the randomized patients, 382 met all eligibility criteria and were assessable for response. The results for tumor response rate are shown in Table 2. The corresponding log odds for the empirical logistic transform are -1.623 and -1.705 for the first row of the table and -1.861 and -1.155 for the second row. Using expression (6) for the approximate variance of an empirical logit, the harmonic mean variance is 0.0741 and this will be used as σ^2/n in the formulas given previously. The model parameters are estimated as $\hat{\beta}_1 = 0.156$ for dose, $\hat{\beta}_2 = 0.0779$ for schedule, and $\hat{\beta}_3 = 0.197$. The standard error of each estimate is $(\sigma^2/4n)^{1/2}$ or 0.1361.

For analysis of these results, we used flat priors for the main effects. The prior distribution for β_3 was taken as normal with mean zero and standard deviation either 0.14 or 0.408. These were the values in which expression (3) gave 0.025 and 0.25 probabilities, respectively, of a qualitative interaction in the specified direction. In this calculation, we used $\Delta = 1.1$, which represents, on the

Table 2
Number of responding patients over number of evaluable patients in study of treatment for advanced ovarian cancer (Eisenhauer et al., 1994)

	Dose	
Duration of infusion	135 mg/m^2 of body surface	175 mg/m^2 of body surface
3 hours	15/91 (16%)	14/91 (15%)
24 hours	14/104 (13%)	23/96 (24%)

(Eisenhauer et al., 1994).

log odds scale, an increase in response rate from 10 to 25%. We take this as the minimum medically important difference.

The posterior mean of β_1, computed from (1), is just the mle $\hat{\beta}_1$.The posterior standard deviation of β_1 is 0.1361, the same as the standard error of the mle. The posterior probability that $\beta_1 > 0$ is 0.87, hence there is not convincing evidence that the average response rate at the 175 mg/m^2 dose is greater than at the 135 mg/m^2 dose. Similarly, the posterior distribution of β_2 has mean 0.078 and standard deviation 0.1361. The posterior probability that $\beta_2 > 0$ is thus 0.72, and there is not convincing evidence that the average response rates differ between the two schedules.

For $\nu = 0.14$, the posterior mean of β_3 is 0.51 times the mle $\hat{\beta}_3$ or 0.101, and the standard deviation is 0.0975. For $\nu = 0.408$, the posterior mean is 0.90 times the mle or 0.177, and the standard deviation is 0.129. Using $\nu = 0.14$, the information about β_3 in the prior is about equal to that in the data. Because the prior has mean zero, the mle for interaction is reduced by about 49% to give the posterior mean of β_3. With $\nu = 0.408$, there is much less information in the prior and the mle is shrunken only by 10%.

The effects of dose for the 3-hour schedule and for the 24-hour schedule are evaluated separately by examining the posterior distributions of $2\beta_1 - 2\beta_3$ and $2\beta_1 + 2\beta_3$, respectively. For $\nu = 0.14$, the means of these distributions are 0.110 and 0.514. These estimates can be contrasted with the corresponding mle's of -0.08 and 0.71. The mle's are more dispersed than the posterior means and, in fact, one mle is negative. The standard deviation of these posterior distributions is $2(0.1361^2 + 0.0975^2)^{1/2}$ or 0.335 in both cases. The 95% highest posterior density intervals for the effect of dose at the two schedules are thus $0.110 \pm 1.96 \times 0.335$ and $0.514 \pm 1.96 \times 0.335$, respectively, or $(-0.54, 0.77)$ for the 3-hour schedule and $(-0.14, 1.17)$ for the 24-hour schedule. As indicated previously, a difference of 1.1 on the log odds scale might be taken as a minimal medically important difference. If this is accepted, then at the 3-hour schedule there is substantial evidence that the effect of dose is not medically significant. For the 24-hour schedule, there is some evidence that the higher dose gives a greater response rate ($\Pr(2\beta_1 + 2\beta_3 \leq 0) = 0.06$), but it is unlikely that the difference is of a medically important magnitude because the minimal medically important difference 1.1 is approximately at the upper boundary of the 95% highest posterior density interval. A more standard analysis based only on the mle's leaves the matter more in doubt. The point estimates are more extreme, the confidence intervals are broader than the highest posterior density intervals, and the interaction is not significant ($p = 0.15$). Results and conclusions for $\nu = 0.408$ are similar, but there is a posterior probability of approximately 0.12 that higher dose is better than lower dose by a medically important amount for the 24-hour schedule.

Consider the planning of this trial using the model described here. We may approximate $\sigma^2 = 1/p(1-p)$, where the probability of response p is expected to be about 0.20; thus, $\sigma^2 = 6.25$. We use flat priors for the main effects and plan the trial to have the 95% posterior density interval for the main effects of treatment be of width $\Delta = 1.1$ corresponding to a minimal medically important difference as described above. Using equation (4), this gives a requirement of approximately $n = 80$ patients in each of the four arms of the trial. One can compute and plot the sample size per arm implied by equation (5) as a function of ν in order to ensure that the width of the 95% posterior interval for the treatment effect for one factor within each level of the other factor would be Δ. The sample size per arm ranges from the value 80 computed from (4) to twice that. The increase is rapid as a function of ν. Even for $\nu = 0.0811$, corresponding to a 5% prior probability of a qualitative interaction, the required sample size is 104 patients per arm. Because a qualitative interaction cannot be viewed as *a priori* impossible, a sample size of at least 104 patients per arm seems reasonable. The investigators originally planned the trial to have 75 patients per arm. Had

they not continued accrual, the data may have been more ambiguous concerning the possibility of important qualitative interactions between schedule and dose.

8. Discussion

The use of a factorial design for a clinical trial is often controversial. Proponents sometimes admit caution if interactions are expected, but also indicate that factorial designs are ideal for studying interactions. Others have questioned how factorial trials can provide meaningful information about interactions when the trials are sized only to detect main effects. Brittain and Wittes (1989) and Peterson and George (1993) have noted that the power for detecting an effect of treatment A is substantially impaired by a negative interaction with treatment B compared to an experiment where all of the patients are randomly allocated either treatment A or placebo, with no use of treatment B. Consequently, factorial designs are often used only when one can assume with confidence that there will be no interactions between the effects of the factors. Some trialists use factorial designs in situations where it is unlikely that more than one factor will be effective, and hence there is less concern about interactions.

When factorial designs are used in clinical trials, the sample size is usually computed based on assuming that there are no interactions. Under this assumption, the average effect of factor A is the same as what has been called the simple effect of treatment A, namely outcome with A alone minus outcome with placebo alone. Factorial trials are usually analyzed by first performing a test of interaction at some level γ. In recognition of the poor power of the conventional interaction test, the level γ is sometimes increased to 0.10 or 0.15. Even with such inflation in level, however, the power of the interaction test is quite limited.

The Bayesian model proposed here provides two advantages over the usual methods used for designing and analyzing factorial trials. First, it encourages the quantification of prior belief about the size of interactions that may exist. Rather than forcing the investigator to adopt one of two extreme positions regarding interactions, it provides for the specification of intermediate positions. The results developed here also permit the investigator to examine the implications of such prior specifications on the interpretability of results, sample size planning, and to compare the factorial design approach to alternatives. One is not restricted to either planning the trial ignoring possible interactions or doubling the sample size in light of the possibility of interactions. The second basic advantage of the Bayesian approach to factorial designs is that the analysis is not based on a preliminary test of interaction having poor power. Ng (1991) has indicated some of the drawbacks of such preliminary tests.

As illustrated in the example described in Section 7, the variance of the posterior distribution of the simple effect of a factor increases rapidly as the variance of the prior distribution of interaction increases from zero. Consequently, in planning a factorial trial where the simple treatment effects are of interest and where the possibility of some interaction cannot be excluded, this Bayesian model suggests that the sample size should be increased by at least 30% compared to a simple two-arm clinical trial for detecting the same size of treatment effect. This is less extreme than doubling the sample size, but is a recommendation quite different from current usage. The 30% figure is somewhat arbitrary. It corresponds to the increase from 80 patients per treatment group to 104 per group in the example to allow for a 5% prior probability of a medically important qualitative interaction between the treatment effects. In other cases, the appropriate increase may be determined from expressions (4) and (5). It may be of interest to develop multistage approaches in which one determines sequentially whether the estimation of main effects will suffice or whether sampling should continue further to estimate the effects of each treatment within levels of the other treatment. There are also trial situations where the main effects, which represent average effects, are more clinically relevant than the simple treatment effects. For example, if factor B represents a palliative treatment that is widely, but not uniformly, used among community physicians, then the effect of a new intervention, averaged over the levels of factor B, may be the most relevant quantity. In such circumstances, the advantage of the factorial design is that it provides a basis for also evaluating the effectiveness of factor B.

Many statisticians are uncomfortable with Bayesian methods because of their dependence of subjective prior distributions. If one uses the model described here with noninformative priors for the main effects (i.e., $\lambda^2, \tau_1^2, \tau_2^2 \to \infty$), then the posterior distributions of the main effects of treatments are normal with mle's as means and the mle variances as the posterior variances. Posterior statements about interactions or the effects of a treatment within a specified level of the other factor do depend on the variance of the prior distribution for the interaction. Although this dependence may be less than comfortable to frequentist statisticians, the usual frequentist analyses correspond to Bayesian analysis with $\nu^2 = 0$ or ∞ and represent extreme positions. The Bayesian

model does provide a basis for sample size planning that leads to an increase in sample size from that conventionally used, and these additional data will serve to lessen the dependence of the analysis on either untestable assumptions or prior specifications.

Résumé

Le plan factoriel 2×2 est souvent recommandé comme constituant une amélioration de l'efficacité des essais cliniques (dans le cas de l'étude de l'influence simultanée de 2 facteurs sur une réponse). La plupart de ces essais sont conçus en supposant qu'il n'y a pas d'interactions des deux facteurs sur la réponse, supposition souvent sujette à caution, toutefois, du fait que de telles interactions, tout à fait envisageables dans les essais cliniques, ne peuvent généralement être testées avec une puissance suffisante en raison des effectifs retenus. C'est pourquoi nous considérons ici l'utilisation de méthodes bayésiennes pour la conception et l'analyse d'essais cliniques en plans factoriels 2×2, cette approche évitant d'avoir à trancher sur l'existence ou non d'interactions. Nous présentons d'abord des résultats exacts dans le cas de plans équilibrés avec une réponse gaussienne. Nous présentons ensuite des approximations, dans le cadre de la modélisation logistique d'une réponse binaire ainsi que pour le modèle des hasards proportionnels qui concerne des données de durée. Les distributions a posteriori qui dérivent de ces approximations sont gaussiennes et ne nécessitent donc pas de calculs compliqués. Nous suggérons enfin des choix possibles pour la distribution a priori.

References

Brittain, E. and Wittes, J. (1989). Factorial designs in clinical trials: The effects of non-compliance and subaddictivity. *Statistics in Medicine* **8,** 161–171.

Buring, J. E. and Hennekens, C. H. (1990). Cost and efficiency in clinical trials: The U.S. physicians' health study. *Statistics in Medicine* **9,** 29–33.

Byar, D. P. and Piantadosi, S. (1985). Factorial designs for randomized clinical trials. *Cancer Treatment Reports* **69,** 1055–1064.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data,* 2nd edition. London: Chapman and Hall.

Eisenhauer, E. A., et al. (1994). European–Canadian randomized trial of Paclitaxel in relapsed ovarian cancer: High-dose versus low-dose and long versus short infusion.*Journal of Clinical Oncology* **12,** 2654–2666.

Fisher, R. A. (1935). *The Design of Experiments.* London: Collier Macmillan.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 1–41.

Miller, R. G. (1981). *Survival Analysis.* New York: Wiley.

Ng, T. (1991). The impact of a preliminary test for interaction in a 2×2 factorial trial. *Proceedings of the Biopharmaceutical Section of the American Statistical Association,* Alexandria, VA, 220–227.

Peterson, B. and George, S. L. (1993). Sample size requirements and length of study for testing interaction in a $2 \times k$ factorial design when time-to-failure is the outcome. *Controlled Clinical Trials* **14,** 511–522.

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: I. Introduction and design. *British Journal of Cancer* **34,** 585–612.

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

Stampfer, M. J., Buring, J. E., Willett, W., Rosner, B., Eberlein, K., and Hennekens, C. H. (1985). The 2×2 factorial design: Its application to a randomized trial of aspirin and carotene in U. S. physicians. *Statistics in Medicine* **4,** 111–116.

Received February 1996; *revised August* 1996; *accepted September* 1996.