BIOMETRICS 56, 213–219
March 2000

Approximate Bayesian Evaluation of Multiple Treatment Effects

Peter F. Thall,[1,*] Richard M. Simon,[2] and Yu Shen[1]

[1]Department of Biostatistics, Box 237, M. D. Anderson Cancer Center,
1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.
[2]Biometric Research Branch, National Cancer Institute,
6130 Executive Boulevard, Room 739, Rockville, Maryland 20852, U.S.A.
* *email:* rex@odin.mdacc.tmc.edu

SUMMARY. We propose an approximate Bayesian method for comparing an experimental treatment to a control based on a randomized clinical trial with multivariate patient outcomes. Overall treatment effect is characterized by a vector of parameters corresponding to effects on the individual patient outcomes. We partition the parameter space into four sets where, respectively, the experimental treatment is superior to the control, the control is superior to the experimental, the two treatments are equivalent, and the treatment effects are discordant. We compute posterior probabilities of the parameter sets by treating an estimator of the parameter vector like a random variable in the Bayesian paradigm. The approximation may be used in any setting where a consistent, asymptotically normal estimator of the parameter vector is available. The method is illustrated by application to a breast cancer data set consisting of multiple time-to-event outcomes with covariates and to count data arising from a cross-classification of response, infection, and treatment in an acute leukemia trial.

KEY WORDS: Bayesian inference; Categorical data; Clinical trials; Multivariate failure times; Survival analysis.

1. Introduction

Patient response to treatment in a clinical trial is often a complex, multivariate phenomenon with components that may differ qualitatively. In cancer trials, where the primary goal is to achieve disease remission and thus extend survival, patients may also experience various adverse events, including transient toxicity or permanent organ damage. When the primary outcome is survival time, treatment effects on disease recurrence time or events characterizing morbidity may be highly relevant in assessing overall treatment effect. In breast cancer trials, a patient may experience some combination of local, regional, distant, or opposite breast recurrence. A treatment may have a different effect on each type of recurrence, possibly including positive effects on some outcomes and negative effects on others. A common method for dealing with multiple event times is to define disease-free-survival (DFS) time as the single outcome for treatment evaluation.

In such settings, summarizing patient outcome by a single variable and characterizing overall treatment effect with a single parameter is a substantive oversimplification that may obfuscate actual effects on specific components of the multivariate patient outcome and thus misrepresent the medical phenomena. We consider it more useful, both scientifically and ethically, to characterize patient outcome and treatment effect as multidimensional objects. This provides a basis for jointly evaluating the risks of specific adverse events and the rates of desired efficacy outcomes. In the case of multiple time-to-event outcomes, evaluating treatment effects on nonfatal events may help to explain the mechanism whereby a given treatment affects overall survival, which may in turn aid in development of new therapies.

In this paper, we propose an approximate Bayesian method for comparative evaluation of two treatments, which we refer to as the experimental (E) and standard (S), based on the results of a randomized clinical trial with multivariate patient outcome. The only requirements to apply the method are that treatment effect can be characterized by a parameter vector $\boldsymbol{\theta}$ having entries corresponding to particular patient outcome variables and that a consistent, approximately normal estimator $\hat{\boldsymbol{\theta}}$ is available. We partition the parameter space into four sets where, respectively, E is superior to S, S is superior to E, the two treatments are equivalent, and the treatment effects are discordant. This partition generalizes the three-set partition of a one-dimensional parameter space based on an indifference region or range of equivalence, as discussed by Spiegelhalter, Freedman, and Parmar (1994) and references cited therein. Under a Bayesian formulation, we obtain an approximately jointly normal posterior for $\boldsymbol{\theta}$ and use this to obtain posterior probabilities of the four parameter sets. The method also accommodates patient prognostic covariates.

The model is established in Section 2, followed by a description of methods for partitioning the parameter space into sets

characterizing the multivariate E-versus-S treatment effect in Section 3. Section 4 presents two applications, the first to multiple time-to-event outcomes with covariates arising from a breast cancer trial and the second to cross-classified count data from an acute leukemia trial. We conclude with a discussion in Section 5.

2. Bayesian Model

The focus of the method is a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ defined so that θ_j characterizes the E-versus-S treatment effect on the jth patient outcome. For example, if $\mathbf{Y} = (Y_1, Y_2)$ is a bivariate binary outcome with Y_1 indicating cancer remission and Y_2 indicating infection, denoting $\pi_{j,t} = \Pr[Y_j = 1 \mid \text{treatment } t]$, the treatment effects may be defined as $\theta_j = g(\pi_{j,\mathrm{E}}) - g(\pi_{j,\mathrm{S}}), j = 1, 2$, for an appropriate transformation g such as the identity, logit, or arcsine square root. Patient covariates $\mathbf{Z} = (Z_1, \ldots, Z_{p-1})$ may be incorporated by defining a treatment indicator $Z_{j,p}$, with $g(\pi_{j,t}) = \beta_{j,1}Z_1 + \cdots + \beta_{j,p-1}Z_{p-1} + \beta_{j,p}Z_{j,p}$, so that the covariate adjusted treatment effects are $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\beta_{1,p}, \beta_{2,p})$. We define the θ_j's so that larger values correspond to superiority of E over S, with $Z_{j,p}$ the indicator of E if the jth outcome is desirable and the indicator of S if it is adverse. The treatment effect vector $\boldsymbol{\theta}$ may be defined similarly for multiple time-to-event outcomes with covariates, as in the breast cancer application in Section 4.1. For discrete Y_j's without covariates, an alternative approach is to start with the multinomial distribution of all possible elementary outcomes and define the treatment effects by summing elementary outcome probabilities as appropriate. We illustrate this approach in Section 4.2, where the method is applied to count data arising from a cross-classification of response, infection, and treatment in a leukemia chemotherapy trial.

For tractability, we require a consistent estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ that is approximately multivariate normal with mean $\boldsymbol{\theta}$ and covariance matrix Σ, denoted $\hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}, \Sigma)$, and we also assume that a consistent estimator of Σ is available. The estimator $\hat{\boldsymbol{\theta}}$ may be obtained via standard maximum likelihood estimation (MLE). Alternatively, depending on the data structure and model assumptions, $\hat{\boldsymbol{\theta}}$ may be obtained via the method of moments (MM) or a generalized estimating equation (GEE) formulation (Liang and Zeger, 1986). For censored time-to-event outcomes, the method of Wei, Lin, and Weissfeld (1989) (subsequently referred to as WLW) may be used.

The distribution theory underlying the method is based on the idea of treating $\hat{\boldsymbol{\theta}}$ as the data vector within a Bayesian formulation under which, *a priori*, $\boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{\mu}, \Omega)$. Utilizing standard MLE, MM, GEE, or WLW distribution theory to establish that $\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \dot{\sim} \mathrm{N}(\boldsymbol{\theta}, \Sigma)$, it follows that, *a posteriori*, $\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}} \dot{\sim} \mathrm{N}(B\mathbf{b}, B)$, where $B^{-1} = \Sigma^{-1} + \Omega^{-1}$ and $\mathbf{b} = \Sigma^{-1}\hat{\boldsymbol{\theta}} + \Omega^{-1}\boldsymbol{\mu}$ (cf., Lindley and Smith, 1972). A desirable aspect of this approximation is that it provides a general basis for Bayesian inference requiring only manipulation of multivariate normal distributions for implementation. This approach has been used by Simon, Dixon, and Freidlin (1996) in the context of Bayesian subset analysis and by Faraggi and Simon (1997) to provide a Bayesian analysis of covariate effects on survival under the Cox model.

3. Treatment Evaluation Criteria

Denote the k-dimensional treatment effect parameter space by Θ. We propose that Θ be partitioned into four sets that characterize the k-dimensional E-versus-S treatment effect vector. Inferences or decisions regarding treatment comparisons may then be based on the posterior probabilities of these sets. In general, we define the partition as follows. Let $\Theta_1 \equiv [\mathrm{E} > \mathrm{S}]$ denote the set of θ where E is superior to S and likewise let $\Theta_2 \equiv [\mathrm{E} < \mathrm{S}]$ be the set where S is superior to E. The equivalence set $\Theta_3 \equiv [\mathrm{E} \sim \mathrm{S}]$ is the set of θ where there is no compelling reason to favor either treatment over the other. The discordance set $\Theta_4 \equiv [\mathrm{E} >< \mathrm{S}]$ is the set where E is superior to S with regard to some effects and inferior to S with regard to others, formally defined as the complement $(\Theta_1 \cup \Theta_2 \cup \Theta_3)^{\mathrm{c}}$ of $\Theta_1 \cup \Theta_2 \cup \Theta_3$. This formulation, and each of the particular constructions of the partition given below, generalize the three-set partition based on the idea of a range of equivalence described by Freedman, Lowe, and Macaskill (1984) in the context of evaluating a one-dimensional treatment effect.

The partition may be formed in a variety of ways, depending on the particular outcomes and trial objectives. To show how this might be done in practice, we describe four different methods for constructing the partition and illustrate each graphically in the two-dimensional case. Each of these approaches begins with a one-dimensional range of equivalence running from a lower limit $\underline{\theta}_j$ to an upper limit $\bar{\theta}_j$ for each treatment effect θ_j. For the jth outcome considered alone, $\theta_j > \bar{\theta}_j$ corresponds to superiority of E over S, $\theta_j < \underline{\theta}_j$ to superiority of S over E, and $\underline{\theta}_j \leq \theta_j \leq \bar{\theta}_j$ to the case where neither treatment is considered superior to the other. These three sets are the one-dimensional versions of Θ_1, Θ_2, and Θ_3, respectively. Discussions of how the one-dimensional cutoffs $\underline{\theta}_j$ and $\bar{\theta}_j$ may be selected in various settings are given by Freedman et al. (1984) and Spiegelhalter et al. (1994). The partition of Θ may be defined in any manner that is medically and scientifically appropriate for the particular application. Ideally, the process of developing a partition should involve medical collaborators or, in nonmedical applications, appropriate subject area specialists.

We first consider the rectangular partition illustrated in Figure 1a. The set where E is superior to S, denoted by $\mathrm{E} > \mathrm{S}$ in the figure, is the set of $\boldsymbol{\theta}$ where all $\theta_j > \underline{\theta}_j$ and at least one $\theta_j > \bar{\theta}_j$. We use the abbreviated set notation $\Theta_1 = [\theta_j > \bar{\theta}_j \; \exists \; j \text{ and } \theta_j > \underline{\theta}_j \; \forall \; j]$. Thus, on Θ_1, E is at least equivalent to S with regard to all effects and superior to S with regard to at least one effect. Similarly, the set where S is superior to E is $\Theta_2 = [\theta_j < \underline{\theta}_j \; \exists \; j \text{ and } \theta_j < \bar{\theta}_j \; \forall \; j]$. The equivalence set, $\Theta_3 = \cap_{j=1}^{k}[\underline{\theta}_j \leq \theta_j \leq \bar{\theta}_j]$, is the k-dimensional rectangle where the treatments are equivalent with respect to each effect considered individually. For each $\boldsymbol{\theta}$ in the discordance set $\Theta_4 = (\Theta_1 \cup \Theta_2 \cup \Theta_3)^{\mathrm{c}}$, there is at least one pair of effects θ_j and θ_l such that $\theta_j > \bar{\theta}_j$ and $\theta_l < \underline{\theta}_l$, i.e., E is superior to S with regard to one effect and S is superior to E with regard to the other. The partition in Figure 1b is defined similarly, with the important difference that $[\theta_1 > \bar{\theta}_1] \subset \Theta_1$ and $[\theta_1 < \underline{\theta}_1] \subset \Theta_2$, so that, if E is either superior or inferior to S with regard to θ_1, the other entries of $\boldsymbol{\theta}$ are irrelevant. This partition is appropriate when θ_1 corresponds to survival, as in the breast cancer application described in Section 4.1.

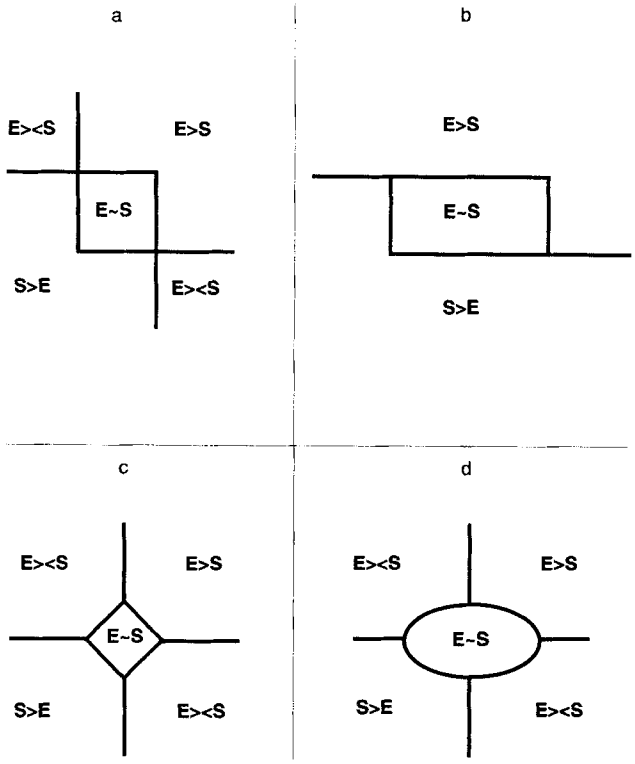

Figure 1. Four possible partitions of a two-dimensional parameter space.

While the definitions of Θ_1 and Θ_2 in Figure 1b imply that $\Theta_4 = \phi$ in the two-dimensional case, Θ_4 is not necessarily empty for $k \geq 3$.

Figure 1c and 1d illustrates partitions based on trade-offs between θ_1 and θ_2. For example, each corner of the diamond in Figure 1c might quantify a trade-off between efficacy and toxicity. Figure 1c is obtained by first defining the $2k$ vectors $\bar{\boldsymbol{\theta}}_j^* = [\theta_j = \bar{\theta}_j$ and $\theta_k = 0, \forall\, k \neq j]$ and $\underline{\boldsymbol{\theta}}_j^* = [\theta_j = \underline{\theta}_j$ and $\theta_k = 0, \forall\, k \neq j]$ for each $j = 1, \ldots, k$. The equivalence set Θ_3 is the convex hull of $\cup_{j=1}^k (\bar{\boldsymbol{\theta}}_j^* \cup \underline{\boldsymbol{\theta}}_j^*)$, represented in the two-dimensional case by the diamond-shaped set. We define Θ_1 to be the open convex hull of $\cup_{j=1}^k [\theta_j > \bar{\theta}_j$ and $\theta_k = 0, \forall\, k \neq j]$. The ideas underlying this definition of Θ_1 are that (1) if E is superior to S at the point $\bar{\boldsymbol{\theta}}_j^*$, then this must also be the case for any $\boldsymbol{\theta}$ such that $\theta_j > \bar{\theta}_j$ and all other $\theta_k = 0$ and (2) if E is superior to S at both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, then this must also be the case for any $\boldsymbol{\theta}$ on the line connecting these two vectors, i.e., Θ_1 should be convex. Similarly, Θ_2 is the open convex hull of $\cup_{j=1}^k [\theta_j < \underline{\theta}_j$ and $\theta_k = 0, \forall\, k \neq j]$, and $\Theta_4 = (\Theta_1 \cup \Theta_2 \cup \Theta_3)^c$. Other approaches to partitioning Θ are possible, such as replacing the polygonal regions in Figure 1a, 1b, and 1c with smoother analogs. Figure 1d illustrates a partition defined by beginning with an equivalence region that is an ellipsoid determined by the points $\bar{\boldsymbol{\theta}}_1^*, \underline{\boldsymbol{\theta}}_1^*, \ldots, \bar{\boldsymbol{\theta}}_k^*, \underline{\boldsymbol{\theta}}_k^*$. The remaining three sets are then defined similarly to the partition in Figure 1c.

4. Applications

In this section we apply the method to two data sets arising from randomized comparative clinical trials. The first consists of multiple time-to-event outcomes with covariates from a breast cancer trial. The second consists of count data from an acute leukemia trial. We chose two data sets with very different structures and used different types of partitions to illustrate the generality of the proposed methodology.

4.1 *Breast Cancer Data*

The first data set arose from a trial conducted by the National Surgical Adjuvant Breast and Bowel Project (NSABP) to compare melphalan + 5-FU (PF) to PF + adriamycin (PAF) for the treatment of women with stage II breast cancer who were considered nonresponsive to tamoxifen (Fisher et al., 1989). The data analyzed here consist of the times to death, local recurrence, and distant recurrence, along with treatment and prognostic covariates.

We first establish some additional notation to deal with multiple time-to-event outcomes with covariates. Let $\mathbf{T} = (T_1, \ldots, T_k)$ denote the vector of k possibly censored time-to-event outcomes, with $\delta_j = 1$ if T_j is the jth event time and $\delta_j = 0$ if it is the independent right-censoring time. Let $(Z_1, \ldots, Z_{p-1})$ denote the patient's covariates, let Z_p be the indicator of the standard treatment, and denote $\mathbf{Z} = (Z_1, \ldots, Z_p)$. Denote the linear term corresponding to T_j by $\xi_j = \boldsymbol{\beta}_j \mathbf{Z}' = \beta_{j,1} Z_1 + \cdots + \beta_{j,p} Z_p$, so that under the usual Cox model (1972), the hazard function is $\lambda_j(t \mid \mathbf{Z}) = \lambda_{j,0}(t)\exp(\xi_j)$, equivalently $\xi_j = \log\{\lambda_j(t \mid \mathbf{Z})/\lambda_{j,0}(t)\}$. Thus, positive values of $\beta_{j,p}$ correspond to superiority of E over S. We employ the method of WLW to obtain the usual estimators of each marginal Cox model, which we denote by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_k)$, and a robust estimate $V_{\boldsymbol{\beta}}$ of the $pk \times pk$ asymptotic covariance matrix of this pk-dimensional vector of estimates. Under the WLW formulation, $\hat{\boldsymbol{\beta}} \sim \text{AN}(\boldsymbol{\beta}, V_{\boldsymbol{\beta}})$; hence, marginally, the k covariate-adjusted treatment effect estimates $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k) = (\hat{\beta}_{1,p}, \ldots, \hat{\beta}_{k,p}) \sim \text{AN}(\boldsymbol{\theta}, \Sigma)$, where Σ is the appropriate $k \times k$ submatrix of $V_{\boldsymbol{\beta}}$.

For the NSABP data, $k = 3$, with $T_1 =$ time to death, $T_2 =$ time to local recurrence, and $T_3 =$ time to distant recurrence. The covariates are $Z_1 =$ tumor size, standardized by subtracting its mean and dividing by its standard deviation, $Z_2 =$ estrogen receptor level similarly standardized, and $Z_3 = 1$ if any positive lymph nodes are present and is 0 otherwise. The treatment indicator $Z_4 = 1$ if the patient received PF, the standard regimen, and is 0 if she received the experimental combination PAF. Thus, $p = 4$, the linear term of the jth outcome is $\xi_j = \beta_{j,1} Z_1 + \beta_{j,2} Z_2 + \beta_{j,3} Z_3 + \beta_{j,4} Z_4, j = 1, 2, 3$, and the covariate-adjusted treatment effect vector is $(\theta_1, \theta_2, \theta_3) = (\beta_{1,4}, \beta_{2,4}, \beta_{3,4})$, with entries corresponding, respectively, to death, local recurrence, and distant recurrence.

Since the experimental regimen PAF involves adjuvant adriamycin, we consider it superior to the standard PF for a given outcome only if it reduces the relative risk of that event by a specified amount, while PF is superior to PAF if PAF increases the relative risk above one by any amount. For death, we use the criterion that PAF must reduce the risk of death relative to PF, $\exp(-\theta_1)$, to 10/11 or less in order to be considered superior to PF, while any value of $\exp(-\theta_1)$ larger than one corresponds to superiority of PF over PAF. Thus, the region of equivalence for the effect θ_1 of PAF relative to PF on death is $[\underline{\theta}_1, \bar{\theta}_1] = [\log(1), \log(1.1)] = [0, .0953]$. Similarly,

Table 1

Parameters of skeptical priors. Each $\mu_j = (\underline{\theta}_j + \bar{\theta}_j)/2$, *each* $\sigma_j(p)$ *is determined so that* $\Pr[\exp(\theta_j) > 4/3] = p$ *for* $p = .05, .10,$ *or* $.25,$ *and* $\rho_{1,3}$ *is determined so that* $\Pr[\exp(\theta_1) > 4/3 \mid \theta_3 = 0] = p/2$, *with* $\rho_{1,2} = \rho_{2,3} = 0$.

		$\underline{\theta}_j$	$\bar{\theta}_j$	μ_j	$\sigma_j(.05)$	$\sigma_j(.10)$	$\sigma_j(.25)$
Death	(θ_1)	0	.0953	.0477	.1459	.1873	.3559
Local recurrence	(θ_2)	0	.2231	.1116	.1071	.1374	.2611
Distant recurrence	(θ_3)	0	.1823	.0912	.1195	.1533	.2914
$\rho_{1,3}$					.2970	.3909	.6472

the respective requirements that PAF must reduce the relative risk of local recurrence to 4/5 and of distant recurrence to 5/6 in order to be considered marginally superior yield equivalence intervals $[\underline{\theta}_2, \bar{\theta}_2] = [0, .2231]$ and $[\underline{\theta}_3, \bar{\theta}_3] = [0, .1823]$. If PAF is either superior or inferior to PF with regard to survival, then θ_2 and θ_3 are irrelevant. We thus constructed a three-dimensional partition that generalizes Figure 1b as follows:

$$\begin{aligned}
\Theta_1 &= [\theta_1 > \bar{\theta}_1] \cup [\theta_2 > \bar{\theta}_2, \theta_1 \geq \underline{\theta}_1, \theta_3 \geq \underline{\theta}_3] \\
&\quad \cup [\theta_3 > \bar{\theta}_3, \theta_1 \geq \underline{\theta}_1, \theta_2 \geq \underline{\theta}_2] \\
\Theta_2 &= [\theta_1 < \underline{\theta}_1] \cup [\theta_2 < \underline{\theta}_2, \theta_1 \leq \bar{\theta}_1, \theta_3 \leq \bar{\theta}_3] \\
&\quad \cup [\theta_3 < \underline{\theta}_3, \theta_1 \leq \bar{\theta}_1, \theta_2 \leq \bar{\theta}_2] \\
\Theta_3 &= [\underline{\theta}_1 \leq \theta_1 \leq \bar{\theta}_1] \cap [\underline{\theta}_2 \leq \theta_2 \leq \bar{\theta}_2] \cap [\underline{\theta}_3 \leq \theta_3 \leq \bar{\theta}_3] \\
\Theta_4 &= (\Theta_1 \cup \Theta_2 \cup \Theta_3)^c.
\end{aligned}$$

As noted earlier, $\Theta_4 \neq \phi$ in this case since it contains, e.g., the set $[\underline{\theta}_1 \leq \theta_1 \leq \bar{\theta}_1] \cap [\theta_2 > \bar{\theta}_2] \cap [\theta_3 < \underline{\theta}_3]$, where θ_1 is in its equivalence interval but θ_2 and θ_3 are discordant. To construct priors, we first considered each θ_j marginally, with the prior mean μ_j of θ_j taken to be the center $(\underline{\theta}_1 + \bar{\theta}_1)/2$ of its equivalence interval to reflect *a priori* equipoise between superiority and inferiority of E compared to S. The standard deviation σ_j was then determined to reflect a given level of skepticism, with several values considered. Because 4/3 is a substantive difference in the relative risk, for given small probability $p = .05, .10,$ or $.25$, we determined σ_j so that $\Pr[\exp(\theta_j) > 4/3] = p$. We assumed that only the effects on death and distant recurrence were correlated, with their correlation $\rho_{1,3}$ determined so that $\Pr[\exp(\theta_1) > 4/3 \mid \theta_3 = 0] = p_c$ for $p_c < p$. In practice, p_c might be elicited from the physicians along with p. Under multivariate normality, $\theta_1 \mid \theta_3 \sim \mathrm{N}(\mu_1 + (\theta_3 - \mu_3)\rho_{1,3}\sigma_1/\sigma_3, \sigma_1^2(1-\rho_{1,3}^2))$; hence, $\rho_{1,3}$ is determined once μ_1, μ_3, σ_1, and σ_3 are given. Since $p_c < p$, we took $p_c = p/2$ to reflect a strong prior association between θ_1 and θ_3. The equivalence intervals and parameters of these three priors are summarized in Table 1.

Table 2 summarizes fits of the marginal Cox model for each outcome, with the standard error of each estimated parameter obtained from the WLW robust covariance matrix estimate. For comparison, the posterior mean and standard deviation of each parameter under the Bayesian formulation with each prior also are given. A 95% posterior credibility interval for each outcome-specific treatment effect, running from the 2.5th to 97.5th percentiles of the posterior, is given in parentheses below the posterior mean and SD. Since each treatment effect in the model is defined as the log relative risk of the event for PF as compared to PAF, the positive-valued estimates in Table 2 correspond to superiority of PAF over PF. The positive-valued estimates of the covariates correspond to increased risk of each event with positive nodes, larger ER, or larger tumor size. The fact that most of the posterior means are closer to zero than the corresponding frequentist parameter estimates reflects shrinkage due to the effect of the prior. The method is rather insensitive to the informativeness of the prior, however, for the priors considered.

Posterior probabilities of the partition sets are presented in Table 3. We considered each skeptical prior and also the prior with all θ_j's i.i.d. $\mathrm{N}(0, 1)$, which is essentially noninformative because $\sigma = 1$ is large relative to $\log(4/3) = .288$. Given that $\Pr[\mathrm{E} > \mathrm{S}] = .881$ with this prior, it is notable that $\Pr[\mathrm{E} > \mathrm{S}] = .834$ under the most skeptical prior where the variances are smallest, a drop of only .047. Thus, the partition probabilities appear to be rather insensitive to the priors considered. The main substantive result is that, after accounting for prognostic covariates and considering the three outcomes jointly using the partition as defined, *a posteriori*, it is highly likely that adriamycin is a desirable adjuvant to PF.

4.2 *The AML Data*

The second data set consists of counts characterizing the rates of complete remission (CR) and infection arising from a chemotherapy trial in acute myelogenous leukemia (Estey et al., 1999). The trial was designed to study the adjuvant effects of all-trans retinoic acid (ATRA) and the growth factor G-CSF, when added to the combination fludarabine + cytosine arabinoside + ifosfamide, in a 2×2 factorial design. For purposes of illustration, we focus on the two treatment groups defined by whether the patient did or did not receive ATRA. Cross-tabulations of the indicators of CR and infection within each of these treatment groups are given in Table 4. A routine examination of these data indicates that the infection rates were very similar (38.3% for the ATRA group vs 38.1% for the no-ATRA group), while the ATRA group had a 55.1% CR rate compared to 46.7% in the no-ATRA group. Although any reasonable treatment comparison based on both of these outcomes should include comparison of the two marginal rates, it is worthwhile to also account for the association within each 2×2 table since the odds ratio is .167 for the ATRA data and .090 for the no-ATRA data. This reflects the fact that, for all patients, the CR rate among patients with infection was much lower than among patients without infection (22.2 vs 68.7%).

Table 2
Frequentist coefficient estimates based on WLW method and approximate Bayesian posterior means and standard deviations

	Frequentist estimates $\hat{\beta}$(SD)	Bayesian estimates $p = .05$ Mean (SD)	$p = .10$ Mean (SD)	$p = .25$ Mean (SD)
Time to Death				
Treatment	.224 (.107)	.191 (.101) (−.007, .389)	.197 (.102) (−.003, .397)	.208 (.104) (.004, .412)
Positive nodes	.546 (.108)	.528 (.107)	.528 (.107)	.527 (.107)
ER	.012 (.060)	.014 (.059)	.014 (.059)	.013 (.059)
Tumor size	.160 (.048)	.160 (.048)	.160 (.048)	.159 (.048)
Time to First Local Recurrence				
Treatment	.336 (.108)	.296 (.101) (.098, .494)	.304 (.103) (.103, .505)	.318 (.105) (.112, .524)
Positive nodes	.648 (.110)	.629 (.108)	.629 (.108)	.629 (.108)
ER	.026 (.054)	.028 (.053)	.027 (.053)	.026 (.053)
Tumor size	.145 (.052)	.145 (.052)	.145 (.052)	.145 (.052)
Time to Distant Recurrence				
Treatment	.210 (.107)	.175 (.101) (.023, .373)	.182 (.102) (−.018, .382)	.193 (.104) (−.011, .397)
Positive nodes	.633 (.109)	.614 (.107)	.614 (.107)	.615 (.107)
ER	.021 (.063)	.022 (.063)	.022 (.063)	.021 (.063)
Tumor size	.133 (.051)	.132 (.051)	.132 (.051)	.132 (.051)

To apply the approximate Bayesian method to these data, let $t = 1$ index the ATRA group and $t = 2$ index the no-ATRA group and let n_1 and n_2 denote their respective sample sizes. For each t, define the random vector $\mathbf{Y}_t = (Y_{t,1}, Y_{t,2}, Y_{t,3})$, where $Y_{t,1}$ is the number of patients with CR and infection, $Y_{t,2}$ is the number with CR and no infection, and $Y_{t,3}$ is the number with infection and no CR. Let $\pi_{t,j}$ denote the probability of the jth outcome in group t with $\boldsymbol{\pi}_t = (\pi_{t,1}, \pi_{t,2}, \pi_{t,3})$. The probability of neither CR nor infection in group t is $\pi_{t,4} = 1 - (\pi_{t,1} + \pi_{t,2} + \pi_{t,3})$, and the corresponding count is $Y_{t,4} = n_t - (Y_{t,1} + Y_{t,2} + Y_{t,3})$. Assuming patients are exchangeable, $\mathbf{Y}_t \mid \boldsymbol{\pi}_t$ is multinomially distributed with parameters $\boldsymbol{\pi}_t$ and n_t, $t = 1, 2$, and $\mathbf{Y}_1, \mathbf{Y}_2$ are independent. Since $\pi_{t,1} + \pi_{t,2}$ is the probability of CR and $\pi_{t,1} + \pi_{t,3}$ is the probability of infection in group t, we may define the ATRA-versus-no–ATRA treatment effects $\theta_{\mathrm{CR}} = g(\pi_{1,1} + \pi_{1,2}) - g(\pi_{2,1} + \pi_{2,2})$ and $\theta_{\mathrm{INF}} = g(\pi_{1,1} + \pi_{1,3}) - g(\pi_{2,1} + \pi_{2,3})$, where $g(\cdot) = \sin^{-1}(\cdot)^{1/2}$ to stabilize the asymptotic variance. Denote $\boldsymbol{\theta} = (\theta_{\mathrm{CR}}, \theta_{\mathrm{INF}})$. Since the usual estimator $\hat{\boldsymbol{\pi}}_t = \mathbf{Y}_t / n_t \sim \mathrm{AN}(\boldsymbol{\pi}_t, V_{\pi,t})$, where $V_{\pi,t}$ has diagonal terms $\pi_{t,j}(1 - \pi_{t,j})/n_t$ and off-diagonal terms $-\pi_{t,j}\pi_{t,k}/n_t$ with $\hat{\boldsymbol{\pi}}_1$ and $\hat{\boldsymbol{\pi}}_2$ independent, it follows by a straightforward application of the delta method that $\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \sim \mathrm{AN}(\boldsymbol{\theta}, \Sigma)$, where Σ is the 2×2 matrix with diagonal terms $(1/4)(n_1^{-1} + n_2^{-1})$ and off-diagonal terms $(1/4)(\rho_1/n_1 + \rho_2/n_2)$, where

$$\rho_t = [\pi_{t,1} - (\pi_{t,1} + \pi_{t,2})(\pi_{t,1} + \pi_{t,3})] \div \{(\pi_{t,1} + \pi_{t,2})(\pi_{t,1} + \pi_{t,3})(1 - \pi_{t,1} - \pi_{t,3}) \times (1 - \pi_{t,1} - \pi_{t,2})\}^{1/2}.$$

Table 3
Posterior probabilities of the four parameter sets for death, local recurrence and distant recurrence from the breast cancer data

Prior	E > S	S > E	E ~ S	E >< S
Skeptical, $p = .05$	.834	.032	.078	.056
Skeptical, $p = .10$	.846	.029	.072	.053
Skeptical, $p = .25$	.866	.025	.055	.054
$\theta_1, \theta_2, \theta_3 \sim$ i.i.d. N(0, 1)	.881	.021	.054	.044

Table 4
Cross-tabulations of CR and infection within each treatment group

CR	Infection: No	Infection: Yes	Totals
ATRA Group			
No	19	29	48
Yes	47	12	59
Totals	66	41	107
No-ATRA Group			
No	22	34	56
Yes	43	6	49
Totals	65	40	105

Thus, the variance stabilizing transformation removes π from the asymptotic variances of $\hat{\theta}_{\mathrm{CR}}$ and $\hat{\theta}_{\mathrm{INF}}$ but not from the formula for the asymptotic correlation.

We may now apply the Bayesian approximation to obtain posterior probabilities of the four two-dimensional parameter sets for assessing the effects of the two treatments on CR and infection. We assume that, *a priori*, θ_{CR} and θ_{INF} are i.i.d. $\mathrm{N}(0, \sigma^2)$, with $\sigma^2 = 2, 1, .5$, or $.05$. For this application, our criterion is that E and S have equivalent marginal probabilities $p_{j,\mathrm{E}}$ and $p_{j,\mathrm{S}}$ if $|p_{j,\mathrm{E}} - p_{j,\mathrm{S}}| \leq .10$. For $p_{j,\mathrm{E}}$ and $p_{j,\mathrm{S}}$ in the range .40 to .60, this produces the equivalence interval $[-.101, .101]$ on the arcsin square root domain, so that $\bar{\theta}_1 = \bar{\theta}_2 = 1.01$ and $\underline{\theta}_1 = \underline{\theta}_2 = -1.01$. Because neither CR nor infection dominates the other outcome in the sense that survival time dominates recurrence time, we employ the symmetric partition illustrated in Figure 1a. This two-dimensional partition takes the following form:

$$\Theta_1 = [\theta_1 > \bar{\theta}_1, \theta_2 \geq \underline{\theta}_2] \cup [\theta_2 > \bar{\theta}_2, \theta_1 \geq \underline{\theta}_1]$$
$$\Theta_2 = [\theta_1 < \underline{\theta}_1, \theta_2 \leq \bar{\theta}_2] \cup [\theta_2 < \underline{\theta}_2, \theta_1 \leq \bar{\theta}_1]$$
$$\Theta_3 = [\underline{\theta}_1 \leq \theta_1 \leq \bar{\theta}_1] \cap [\underline{\theta}_2 \leq \theta_2 \leq \bar{\theta}_2]$$
$$\Theta_4 = (\Theta_1 \cup \Theta_2 \cup \Theta_3)^{\mathrm{c}}.$$

For $.5 \leq \sigma^2 \leq 2$, the approximate bivariate normal posterior of $[(\theta_{\mathrm{CR}}, \theta_{\mathrm{INF}}) \mid \hat{\boldsymbol{\theta}}]$ has mean $(.084, .002)$, common variance .0047, and $\mathrm{cov}(\theta_{\mathrm{CR}}, \theta_{\mathrm{INF}} \mid \hat{\boldsymbol{\theta}}) = -.0021$. Under the more skeptical prior with $\sigma^2 = .05$, these values change very slightly to mean $(.077, .005)$, variance .0042, and covariance $-.0018$. For $\sigma^2 \geq .5$, the posterior means are numerically identical to the corresponding MLEs. Importantly, although the prior correlation between θ_{CR} and θ_{INF} is assumed to be zero, the posterior correlation is between $-.45$ and $-.47$ for the above values of σ^2. This reflects, in terms of the posterior bivariate normal correlation, the strong association noted earlier between the CR and infection counts in terms of the usual frequentist estimates of within-group odds ratios.

Under priors with $\sigma^2 = 2, 1$, or $.5$, *a posteriori*, $\Pr[\mathrm{E} > \mathrm{S}] = \Pr[\text{ATRA is superior}] = .404$ to $.410$, $\Pr[\mathrm{S} > \mathrm{E}] = \Pr[\text{no ATRA is superior}] = .068$ to $.069$, $\Pr[\text{ATRA is equivalent to no ATRA}] = .514$ to $.520$, and the discordance set has probability .008. Under the more skeptical prior with $\sigma^2 = .05$, these four probabilities are .365, .065, .564, and .007. An essential point in interpreting and using these probabilities inferentially is that they pertain to the joint posterior distribution of the treatment effects on CR and infection. Thus, they account for both the desirable and undesirable effects of treatment simultaneously. This assignment of posterior probabilities to these two-dimensional parameter sets thus provides an interpretation to these data that is simply not available using the more conventional frequentist methods of reporting the MLEs or performing separate tests comparing the treatments in terms of either CR or infection. As was the case with the breast cancer data, for these count data, the method is remarkably insensitive to the informativeness of the assumed prior.

5. Discussion

We have proposed a Bayesian method for accommodating multiple endpoints and multiple covariates in the monitoring and analysis of randomized clinical trials. Although frequentist methods have been developed for multiple endpoint trials, they are not widely used. Most often, formal inference is based on a single primary endpoint, such as DFS in cancer trials, which is a composite of two or more outcomes. Such composite endpoints may mask treatment effects on individual endpoints. Moreover, the use of composite endpoints as a basis for claiming treatment differences is sometimes controversial. While dimension reduction may be motivated formally by the notion of sufficiency, the use of DFS is difficult to rationalize on that basis. The Bayesian approach described here provides a flexible framework for evaluating treatment effects on multiple endpoints without requiring either a composite endpoint or a hierarchy of importance of the endpoints. It also avoids a penalty for conducting multiple comparisons to make simultaneous inferences about multiple relevant endpoints.

With frequentist methods, there is controversy as to what point or interval estimates of treatment effect are appropriate if a trial is terminated early since MLEs may be quite biased in that circumstance. Although several methods for using the frequentist sequential monitoring boundary to define an adjusted estimate have been proposed, the uncorrected MLE is generally used in medical publications. The situation is even more complex with multiple endpoints and frequentist methods. Under the Bayesian approach described here, inference is straightforward. The mode of the posterior distribution is the most appropriate point estimate and the highest posterior density intervals are the most appropriate interval estimates. Since the posterior distribution is multivariate normal, these estimates are very easy to calculate. Moreover, these estimates are not modified by the use of the decision regions for interim monitoring. Although the estimates for one endpoint will be influenced by the results for other endpoints due to their inherent posterior association, an aspect of the method that we regard as an advantage, the estimates will not be directly dependent on the number of endpoints. A limitation of the methods described here is that they depend on large-sample approximations and thus may be of limited accuracy at the time of an early interim analysis. This is also true of analogous frequentist methods, however.

Our use of a normal prior for the parameters could be generalized. For example, rather than specifying correlations between the treatment effects, one could place a hyper-prior on those correlations, as in Dixon and Simon (1991, 1992). While this would complicate computation of the posterior, Markov chain Monte Carlo or importance sampling methods (Gilks, Richardson, and Spiegelhalter, 1996) could be applied.

The Bayesian framework described here also accommodates subset analysis as described by Simon et al. (1996). This is easily done by extending the model to include treatment–covariate interactions for binary covariates and employing a skeptical prior for the size of the regression coefficients of the interaction effects. If this prior distribution is normal, as in Simon et al. (1996), then the posterior distribution remains multivariate normal. With multiple endpoints, there would be treatment–covariate interactions for each endpoint of interest. Simon et al. (1996) provide details on how to specify skeptical priors for these interaction effects by eliciting conditional probabilities of the medically important treatment effects. Our method also may be generalized to accommodate trials with more than two treatment arms, with the definition of the decision regions reflecting the relationships among the treatments.

Acknowledgements

The authors thank John Bryant and the NSABP for providing data on the clinical trial B-11.

Résumé

Nous proposons une méthode bayésienne approchée pour comparer un traitement expérimental à un traitement de référence basée sur un essai clinique randomisé avec résultat multivarié pour les patients. L'effet global du traitement est caractérisé par un vecteur de paramètres correspondant aux effets sur les résultats individuels des patients. Nous partitionons l'espace des paramètres en quatre ensembles dans lesquels, respectivement, le traitement expérimental est supérieur au témoin, le témoin est supérieur à l'expérimental, les deux traitements sont équivalents, et les effets des traitements sont discordants. Nous calculons les probabilités *a posteriori* de ces ensembles en traitant un estimateur du vecteur de paramètres comme une variable aléatoire selon le paradigme bayésien. L'approximation peut être utilisée dans tout contexte où l'on dispose d'un estimateur convergent et asymptotiquement normal du vecteur de paramètres. La méthode est appliquée à des données de cancer du sein consistant en délais de survenue multiples avec covariables, et à des données de dénombrement dans une classification croisée en réponse, infection et traitement tirées d'un essai dans la leucémie aiguë.

References

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

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

Dixon, D. O. and Simon, R. (1992). Bayesian subset analysis in a colorectal cancer clinical trial. *Statistics in Medicine* **11,** 13–22.

Estey, E. H., Thall, P. F., Pierce, S., Cortes, J., Beran, M., Kantarjian, H., Keating, M., Andreeff, M., and Freireich, E. (1999). Randomized phase II study of fludarabine + cytosine arabinoside + idarubicin ± all-trans retinoic acid ± G-CSF in poor prognosis newly diagnosed non-APL, AML and MDS. *Blood,* in press.

Faraggi, D. and Simon, R. (1997). Large sample Bayesian inference on the parameters of the proportional hazard model. *Statistics in Medicine* **16,** 2573–2585.

Fisher, B., Redmond, C., Wickerham, D. L., et al. (1989). Doxorubicin-containing regimens for the treatment of stage II breast cancer: The National Surgical Adjuvant Breast and Bowel Project Experience *Journal of Clinical Oncology* **7,** 572–582.

Freedman, L. S. and Spiegelhalter, D. J. (1992). Application of Bayesian statistics to decision making during a clinical trial. *Statistics in Medicine* **11,** 23–35.

Freedman, L. S., Lowe, M., and Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. *Biometrics* **40,** 575–586.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* New York: Chapman and Hall.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

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.

Simon, R., Dixon, D. O., and Freidlin, B. (1996). Bayesian subset analysis of a clinical trial for the treatment of HIV infections. In *Bayesian Biostatistics,* D. A. Berry and D. Stangl (eds), 555–576. New York: Marcel Dekker.

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

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

Received August 1998. *Revised March* 1999.
Accepted May 1999.