Bone Marrow Transplantation (2004) 33, 1191–1199
© 2004 Nature Publishing Group All rights reserved 0268-3369/04 $30.00
www.nature.com/bmt

Conditioning regimens

Comparison of 100-day mortality rates associated with i.v. busulfan and cyclophosphamide *vs* other preparative regimens in allogeneic bone marrow transplantation for chronic myelogenous leukemia: Bayesian sensitivity analyses of confounded treatment and center effects

PF Thall[1], RE Champlin[2] and BS Andersson[2]

[1]*Department of Biostatistics, University of Texas, MD Anderson Cancer Center, Houston, TX, USA; and* [2]*Department of Blood and Marrow Transplantation, University of Texas, MD Anderson Cancer Center, Houston, TX, USA*

Summary:

We evaluated the 100-day mortality rates associated with busulfan-based myeloablative conditioning regimens based on data from 1812 chronic myelogenous leukemia patients who underwent allogeneic blood or marrow transplantation (allotx). In all, 47 patients received intravenous (i.v.) busulfan and cyclophosphamide (i.v.BuCy2) with allotx at MD Anderson Cancer Center (MDACC) during 1995–1999. The remaining 1765 patients, whose data were supplied by the International Bone Marrow Transplant Registry (IBMTR), received alternative preparative regimens, primarily Cy-total body irradiation (~45%) or oral BuCy (~35%) during 1997–1998. As patients were not randomized between conditioning regimens, the i.v.BuCy2-*versus*-alternative treatment effect is confounded with a possible center effect due to nontreatment differences associated with factors differing between MDACC and the IBMTR centers. Additional complications are that the i.v.BuCy2-MDACC patients all survived 100 days, and three prognostic subgroups were included. Bayesian sensitivity analyses were performed to assess treatment effect on the probability of 100-day mortality, over a range of possible MDACC-*versus*-IBMTR center effects. For these patients, the posterior probability that i.v.BuCy2 was superior to alternative conditioning regimens ranges from 0.54 to 0.99, depending on prognosis and the magnitude of the assumed center effect.

Bone Marrow Transplantation (2004) **33,** 1191–1199.
doi:10.1038/sj.bmt.1704461
Published online 3 May 2004
Keywords: Bayesian statistics; busulfan; stem cell transplantation; cyclophosphamide

Allogeneic hematopoietic stem cell transplantation (allotx) is a well-established curative therapy for chronic myelogenous leukemia (CML) patients.[1] The most commonly used pretransplant conditioning treatment is total body irradiation (TBI) combined with intravenous (i.v.) cyclophosphamide (Cy).[2] The delivery of TBI is very precise, but its use is burdened with a variety of complications, for example, cataracts, secondary tumors, and degraded intellectual function.[3–12] As an alternative, high-dose oral busulfan (Bu) in combination with Cy was introduced,[13,14] and with a subsequently modified Cy dose given over 2 days ('BuCy2'),[15] the regimen has become widely accepted; it is now the most commonly used non-TBI-based pretransplant treatment.[2] A recent randomized study suggested that long-term outcome after allotx for CML was improved when BuCy2 rather than Cy-TBI was used as the conditioning regimen.[16,17] This is contrary to available data in the International Bone Marrow Transplant Registry[2] (IBMTR), however, where differences in outcome cannot be attributed to conditioning regimen, but only to the clinical stage of CML and whether matched related rather than matched unrelated donors have been used. The oral BuCy2 regimen is generally well tolerated, but several investigators have associated systemic Bu exposure with serious hepatic and neurologic toxicity.[18–24] In addition, the hepatic first-pass Bu exposure after oral drug administration has been suggested to contribute to veno-occlusive disease.[25]

Another problem with oral BuCy2 is that the typically unpredictable and erratic intestinal absorption of Bu contributes to wide interpatient variation in bioavailability, as quantified by the area under plasma concentration *vs* time curve (AUC). Variability in AUC as high as 10-fold or more has been observed between patients given the same oral Bu dose.[26,27] In addition, within-patient variability in systemic exposure, as measured repeatedly after a fixed dose given every 6 h, is also substantial, and may be as high as two- to threefold. This variability greatly contributes to uncertainty with regard to actual delivered dose.[28] Contrary to this, the various factors that influence intestinal drug absorption are completely circumvented by i.v. drug administration, and (i.v. delivery therefore yields a) 'desired concentration of a drug in blood…with an accuracy and immediacy not possible with any other procedure', as was eloquently

Correspondence: Dr PF Thall, Department of Biostatistics, Box 447, University of Texas, MD, Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA; E-mail: rex@mdanderson.org
Received 13 August 2003; accepted 2 December 2003
Published online 3 May 2004

pointed out by Benet and Sheiner,[29] with regard to i.v. drug formulations. As precise, predictable dose exposure is very important in pretransplant conditioning therapy, we developed a pharmaceutically acceptable i.v.Bu formation.[27,30,31] We have demonstrated that this novel formulation yields highly reproducible pharmacokinetics (PK) with minimal within-patient variability in PK parameters and a patient-to-patient variability in AUC that was reduced to two-fold, approximately.[31] Scientifically, the ideal approach to comparing the efficacy of these preparative regimens would be to embark on a confirmatory phase III study, with patients randomized between i.v.BuCy2 and a chosen 'standard' regimen, stratifying patients for clinical stage and other prognostic factors. However, at present there are some severe drawbacks with this approach. Such a randomized study would require the participation of several hundred patients and would take a long time to complete. A major problem with initiating such a study now arises from the fact that the current method for choosing an i.v.BuCy2 dose is to optimize it for individual patients using preliminary patient-specific PK information.[32] This study strongly indicated that individualized i.v.Bu dosing based on PK parameters is superior to a fixed-dose Bu-based regimen. As this within-patient dose optimization method is still being refined,[32] it seems inappropriate to begin a long-term study of i.v.BuCy2 now. This is because it would be necessary to either continue to refine the method throughout the phase III trial or, alternatively, fix the dose optimization procedure to be a particular method that might well prove to be suboptimal midway through the trial. The former approach would introduce heterogeneity in the 'i.v.BuCy2' regimen and possibly create a drift over time in the treatment effect. The latter approach would produce results that likely would be outdated before completion of the study. Furthermore, continuing to randomize patients to such a study after a substantive improvement in the within-patient dose optimization method for i.v.BuCy2 would be undesirable for many physicians.

After treating and evaluating a series of 47 patients with i.v.BuCy2, we observed the striking result that none of these patients died within the first 100 days post transplant. This was very different from the concomitant experience with alternative (Alt) preparative regimens, as documented by the IBMTR. Consequently, we undertook a comparison of 100-day mortality between these 47 MDACC patients and 1765 IBMTR CML allogeneic transplant patients treated during a comparable time period, 1997–1998. All of the IBMTR patients received Alt preparative regimens, primarily Cy-TBI (~45%) or oral BuCy (~35%). Unfortunately, this comparison was complicated by several scientific issues. First, patients were not randomized between preparative regimens. Second, all patients who received i.v.BuCy2 were treated at MDACC while all patients receiving Alt regimens were treated at other participating IBMTR centers. Consequently, the i.v.BuCy2-versus-Alt treatment effect is confounded by other effects due to differences between the MDACC patients and the IBMTR patients not pertaining to their preparative regimens. Such differences may have arisen from variability in patient characteristics, supportive care, physicians' experience and skills, time periods of treatment, or other variables. While the effects of individual patient prognostic covariates could be accounted for by a statistical regression analysis, such data were not available for the 1765 IBMTR patients. We will refer to the composite MDACC-versus-IBMTR difference, arising from effects other than pretransplantation conditioning treatment, as the 'center effect'. In sharp contrast with the fact that none of the 47 i.v.BuCy2-MDACC patients died before day 100, the 100-day death rates of the 1765 Alt-IBMTR patients varied from 18 to 30%, depending on disease stage. Consequently, the statistical problem is to compare treatment effects between two sets of binomial samples while accounting for treatment-center confounding, prognostic subgroup, and the fact that all of the samples in one set have 0 events.

Despite the presence of treatment-center confounding, the currently available data may provide useful insights into what the actual comparative effects of the i.v.BuCy2 and Alt preparative regimens may be. We addressed the problem of comparing treatments in the presence of treatment-center confounding here by first assuming that, within each prognostic subgroup, the observed difference was the sum of a hypothetical treatment effect and a center effect. A Bayesian model,[33,34] under which each of the effects of interest is considered to be a random quantity, was assumed. We performed a sensitivity analysis under which the putative center effect was varied over a range of values, accounting for anywhere from none to all of the observed difference, and the resulting treatment effect was computed. This is similar to the method used by Estey *et al*,[35] to evaluate possible treatment effects in the presence of treatment-trial confounding. While our Bayesian sensitivity analysis is not a substitute for a randomized trial, it can be used as a means to assess what the unknown treatment effects are likely to be. This may be used, in turn, as a basis for deciding whether to proceed with a prospective randomized trial, as well as making therapeutic decisions in the clinic during the time that the delivery of i.v.Bu is optimized, and also subsequently while future long-term randomized studies of i.v.Bu-based conditioning therapy *vs* Alt regimen(s) are ongoing.

Patients and treatments

We treated 47 consecutive CML patients undergoing allotx for various stages of their disease at MDACC from July 1996 to October 1999. Of these patients, 17 were in chronic phase (CP), 25 were in accelerated phase (AP), and 5 were in blast crisis (BC) at the time of allotx. In all, 29 (62%) were male and 18 (38%) were female, with median age 40 years (range, 19–64 years). The median time from diagnosis to transplant, in months, was 17 for CP (range, 2–244 months), nine for AP (range, 2–77 months), and 18 for BC (range, 4–111 months). The pretransplant conditioning regimen consisted of i.v.Bu at 0.8 mg/kg body weight over 2 h every 6 h for 16 doses, followed by two daily doses of Cy at 60 mg/kg ('i.v.BuCy2').[31] The drugs were used as a fixed-dose i.v.BuCy2 regimen without PK-guided individualized dose adjustment. In total, 30 patients received bone marrow and 17 received peripheral blood progenitor cells from HLA-matched related donors. The graft-versus-host disease prophylaxis was tacrolimus and minidose methotrexate.

The comparison group consisted of 1765 IBMTR CML transplant patients who received Alt preparative regimens. Of these patients, 1344 (76%) were in CP, 335 (19%) were in AP, and 86 (5%) were in BC. The IBMTR patients received Alt preparative regimens, most commonly Cy-TBI or oral BuCy2.[2]

Statistical methods

Conventional comparisons

The 100-day mortality data of the i.v.BuCy2-MDACC patients and Alt-IBMTR patients, within each CML prognostic subgroup and also overall for the combined prognostic groups, were compared using Fisher's exact test.

Bayesian model

The basic principles of the Bayesian paradigm and properties of the binomial and beta probability models used here are described in the appendix. To establish the Bayesian probability model, for simplicity, we temporarily limit attention to patients in a single prognosis and treatment-center combination, denoted by π, the probability of a given patient in the subgroup dying within 100 days. Here, we assume that, *a priori*, before any data are observed, $\pi \sim \text{beta}[\frac{1}{2}, \frac{1}{2}]$. This prior distribution has mean value $\frac{1}{2}$ and variance 0.125, hence standard deviation (s.d.) 0.354; it contains as much information as knowing the 100-day mortality outcome of one allotx patient. The purpose of this prior assumption is to avoid introducing personal bias or artificial information into the analyses. Once the 100-day mortality data have been observed, say X deaths and Y survivals in $N = X + Y$ patients, the posterior distribution of π is $\text{beta}[\frac{1}{2} + X, \frac{1}{2} + Y]$. Observing these data thus changes the mean of π from the prior value $\frac{1}{2}$ to the posterior value $\mu = (\frac{1}{2} + X)/(N + 1)$, and it also changes the variance of π from the prior value 0.125 to the posterior value $\mu(1-\mu)/(N+2)$. From the appendix, the posterior mean may be expressed as the weighted average, $\mu = \frac{1}{2}\{1/(N+1)\} + (X/N)\{N/(N+1)\}$, of the prior mean, $\frac{1}{2}$, and the empirical mean, X/N. The respective weights, $1/(N+1)$ and $N/(N+1)$, are proportional to the prior sample size, 1, and the actual sample size, N. Since no 100-day deaths were recorded ($X = 0$) in each i.v.BuCy2-MDACC prognostic subgroup, the posterior mean is $\frac{1}{2}/(N+1)$, a value only slightly above 0 in each case.

Bayesian comparisons

Temporarily focus attention on one prognostic subgroup, and denote the 100-day mortality probabilities by π_1 for the i.v.BuCy2-MDACC patients and π_2 for the Alt-IBMTR patients. We will compare the Alt-IBMTR and i.v.BuCy2-MDACC groups in terms of the posterior distributions of their 100-day mortality probabilities, π_2 and π_1. A useful method is to graph the posteriors of π_1 and π_2 and compare them visually. A single statistic that summarizes the difference between π_1 and π_2 is $\Pr(\pi_2 > \pi_1 | \text{Data})$, the posterior probability that the Alt-IBMTR patients had a higher 100-day mortality rate than the i.v.BuCy2-MDACC patients, given the observed data in the two subgroups. If the posteriors of π_1 and π_2 were identical, then this probability would be $\frac{1}{2}$. Values greater (less) than $\frac{1}{2}$ correspond to a lower (higher) 100-day death rate in the i.v.BuCy2-MDACC patients. To obtain an overall comparison based on the combined prognostic subgroups, we computed a weighted average of the three posterior probabilities from the CP, AP, and BC subgroups, with the weights proportional to the three sample sizes. From Table 1, the proportions of patients in the three prognostic subgroups are $(17 + 1344)/1812 = 0.75$ for CP, $(25 + 335)/1812 = 0.20$ for AP, and $(5 + 86)/1812 = 0.05$ for BC. Consequently, the weighted average probability is $0.75 \times \Pr(\pi_{2,\text{CP}} > \pi_{1,\text{CP}} | \text{CP Data}) + 0.20 \times \Pr(\pi_{2,\text{AP}} > \pi_{1,\text{AP}} | \text{AP Data}) + 0.05 \times \Pr(\pi_{2,\text{BC}} > \pi_{1,\text{BC}} | \text{BC Data})$. The appendix provides an Splus program that computes probabilities of the form $\Pr(\pi_2 > \pi_1)$ when π_1 and π_2 each have a beta distribution.

Bayesian sensitivity analyses

Again, for simplicity, temporarily consider a single prognostic subgroup. All of the distributions discussed here are posteriors, that is, conditional on the observed data. Since $\Pr(\pi_2 > \pi_1 | \text{Data}) = \Pr(\pi_2 - \pi_1 > 0 | \text{Data})$, we will focus on the difference of the two 100-day mortality probabilities, $\pi_2 - \pi_1$, which is the comparative effect due to Alt-IBMTR *vs* i.v.BuCy2-MDACC. The main difficulty is that $\pi_2 - \pi_1$ is not the effect of interest, namely the treatment effect, Alt-*vs*

Table 1 Survival (100 days) by preparative regimen, center, and CML prognostic subgroup

	# Deaths within 100 days/# patients (%)		*Fisher's exact test P-value*	*Probability i.v.BuCy2-MDACC has lower 100-day mortality than Alt-IBMTR*
Prognostic subgroup	i.v.BuCy2, MDACC	Alt preparative regimens, IBMTR		
CP	0/17 (0)	242/1344 (18)	0.055	0.991
AP	0/25 (0)	84/335 (25)	0.002	>0.999
BC	0/5 (0)	26/86 (30)	0.316	0.945
Combined	0/47 (0)	352/1765 (20)	<0.001	0.990

i.v.BuCy2. Rather, $\pi_2-\pi_1$ is the effect of Alt preparative regimens confounded with IBMTR *vs* i.v.BuCy2 confounded with MDACC. The objective of the sensitivity analyses will be to assess, based on the available data and making some additional model assumptions, what the treatment effect may be. This will be done first within each prognostic subgroup, and then combining the three subgroups by using weighted averages of the form given above. The sensitivity analyses are based on the key assumption that the confounded effect, $\pi_2-\pi_1$, equals the sum of two hypothetical components, {treatment effect} + {center effect}. Given this, varying the unknown proportion of $\pi_2-\pi_1$ that is due to center effect from 0 to 1 provides a basis for assessing the corresponding putative treatment effects.

In terms of the posteriors on π_1 and π_2, there is a reduction in 100-day mortality going from the Alt-IBMTR patients to the i.v.BuCy2-MDACC patients within each prognostic subgroup. For example, in the CP subgroup, the posterior means are $(\frac{1}{2}+242)/(1344+1)=0.180$ for Alt-IBMTR and $(\frac{1}{2}+0)/(17+1)=0.028$ for i.v.BuCy2-MDACC, and $\Pr(\pi_2-\pi_1>0|\text{Data})=0.991$ (Table 1). Thus, if there is a center effect, it almost certainly constitutes an advantage due to being treated at MDACC compared to the centers contributing to the IBMTR data.

Fix a value of p between 0 and 1. We will assume that the proportion p of $\pi_2-\pi_1$ is due to center and the remaining $1-p$ is due to treatment. Denote by $\pi_1(p)$ the corresponding hypothetical 100-day mortality probability of an i.v.BuCy2 patient, if center effect could be completely removed. Denote the posterior means of π_1 and π_2 by μ_1 and μ_2, and let N_1 be the posterior effective sample size of π_1, so that π_1 $|\text{Data}\sim\text{beta}[\mu_1N_1,(1-\mu_1)N_1]$. Define $\mu_1(p)=(1-p)\mu_1+p\mu_2$, a weighted average of the means of π_1 and π_2. Our additional model assumption, which will provide a basis for the sensitivity analyses, is that $\pi_1(p)\sim\text{beta}[\mu_1(p)N_1, \{1-\mu_1(p)\}N_1]$. That is, we assume that, given the data, the hypothetical 100-day mortality probability of an i.v.BuCy2 patient, with center effects removed, follows a beta distribution with mean $\mu_1(p)=(1-p)\mu_1+p\mu_2$ and effective sample size N_1, the same as that of π_1. These assumptions ensure that, while the mean of the hypothetical probability $\pi_1(p)$ equals an average of the means of π_1 and π_2, the amount of information in the posterior of $\pi_1(p)$ is the same as the amount of information in the posterior of π_1. Since $\pi_2-\pi_1=\{\pi_2-\pi_1(p)\}+\{\pi_1(p)-\pi_1\}$, the hypothetical Alt-versus-i.v.BuCy2 treatment effect with the confounding center effect removed is $\pi_2-\pi_1(p)$, and the hypothetical center effect is $\pi_1(p)-\pi_1$. If $p=0$, then there is no center effect, $\pi_1(p)=\pi_1(0)$ has the same distribution as π_1, $\Pr(\pi_2>\pi_1(p)|\text{Data})=\Pr(\pi_2>\pi_1|\text{Data})$, and consequently the observed difference $\pi_2-\pi_1$ is the true treatment effect. If $p=\frac{1}{2}$, then $\pi_1(p)=\pi_1(\frac{1}{2})$ has mean $\frac{1}{2}\mu_1+\frac{1}{2}\mu_2$, and on average half of the observed difference is due to treatment effect and half to center effect. If $p=1$, then $\pi_1(p)=\pi_1(1)$ has the same mean as π_2 (but a smaller variance), all of the observed effect is due to center, there is no treatment effect, and $\Pr(\pi_2>\pi_1(p)|\text{Data})$ is approximately $\frac{1}{2}$. This probability is not exactly $\frac{1}{2}$ due to the difference in variances. Given this structure, the sensitivity analysis consists of evaluating $\Pr(\pi_2>\pi_1(\text{p})|\text{Data})$ as p is varied between 0 and 1. We did this for each prognostic subgroup, and also for the weighted average of the three subgroups.

Results

Table 1 summarizes the raw data, *P*-values of the Fisher's exact tests comparing the observed mortality rates in each prognostic group and overall, and the corresponding Bayesian posterior probabilities that one rate is lower than the other. Within each row of Table 1, both the Fisher's exact test *P*-value and the Bayesian comparison reflect the difference between the confounded treatment-center effects, namely i.v.BuCy2-MDACC-versus-Alt-IBMTR, rather than the i.v.BuCy2-versus-Alt treatment effect. Thus, these comparisons are potentially misleading in that they ignore treatment-center confounding and compare the combined effect of receiving i.v.BuCy2 at MDACC to the combined effect of receiving an Alt preparative regimen at an IBMTR center. The Fisher's exact tests show that, using the conventional cutoffs 0.10, 0.05 and 0.01 for the test *P*-values to quantify 'marginal significance', 'significance' and 'high significance', respectively, the difference is marginally significant in CP patients, highly significant in AP patients, and not significant in BC patients. If one ignores CML stage, then an overall test has *P*-value <0.001, which is highly significant.

From a Bayesian perspective, among the i.v.BuCy2-MDACC patients in CP, observing that 0/17 died within 100 days (Table 1) gives a $\text{beta}(\frac{1}{2}+0, \frac{1}{2}+17)$ posterior for π, which has mean 0.0278 and s.d. 0.0377. This posterior mean does not equal 0 because it equals the weighted average $(1/18)\ \frac{1}{2}+(17/18)\ 0$ of the prior mean $\frac{1}{2}$ and the empirical mean 0. This illustrates a general logical advantage of the Bayesian approach; use of the posterior mean as a point estimator of the probability π improves on the nonsensical empirical estimate 0/17. The empirical estimate says that, based on 17 observations, death within 100 days is impossible. Observing that 242/1344 Alt-IBMTR CP patients died within 100 days gives a $\text{beta}(\frac{1}{2}+242,\ \frac{1}{2}+1102)$ posterior. This has mean 0.1803 and s.d. 0.0105, reflecting both the much higher observed 100-day mortality rate and the much larger sample size, hence the much smaller s.d. The posteriors for the other prognostic groups, and for the combined groups, are computed analogously. These posteriors are graphed in Figure 1 for each prognostic subgroup and for the combined subgroups. Each of the two posteriors in Figure 1d for the combined subgroups is obtained as a weighted average of three individual subgroup posteriors. The numerical values on the vertical axes in Figure 1 reflect the fact that the total area under each curve must equal 1, since it is a probability distribution.

The comparisons provided by Figure 1 clearly indicate that, within each prognostic subgroup and overall, the i.v.BuCy2-MDACC patients had a much smaller probability of 100-day mortality than the Alt-IBMTR patients. The last column of Table 1 gives the posterior probability $\Pr(\pi_2>\pi_1|\text{Data})$ corresponding to each of these visual comparisons.

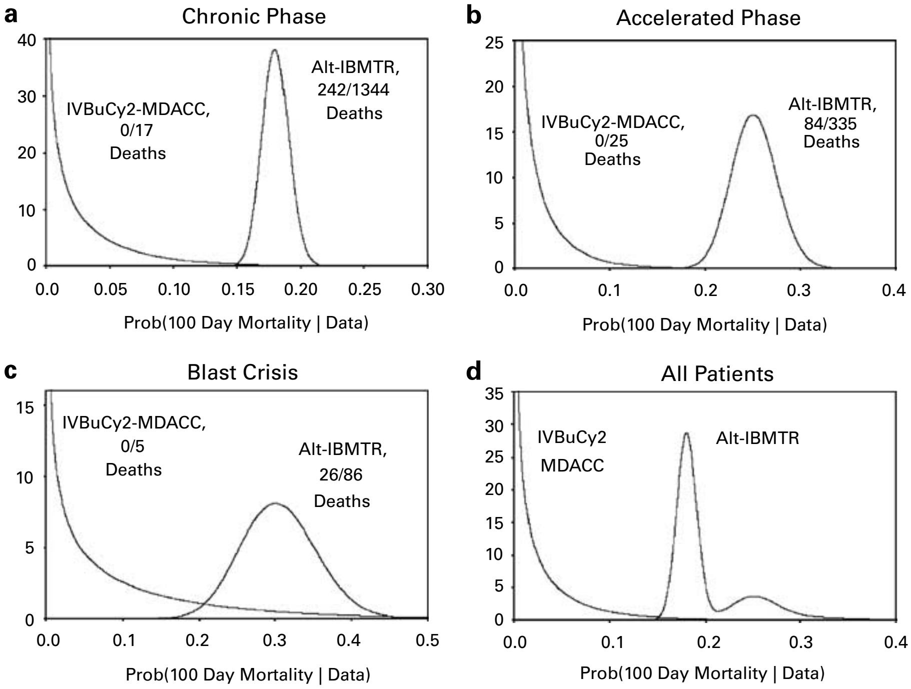

Figure 1 Posterior distributions of the probability of 100-day mortality for the i.v.BuCy2-MDACC patients and the Alt-IBMTR patients, within each CML prognostic group, (**a**–**c**), and for the combined groups (**d**).

Table 2 summarizes the results of four similar Bayesian sensitivity analyses, one within each prognostic subgroup and one for the combined subgroups. In each analysis, we assume that a varying proportion of the observed effect is due to an MDACC-versus-IBMTR center effect. As a basis for comparison, the first column of the table, labeled '0%', gives the probability $\Pr(\pi_2 > \pi_1|\text{Data})$, within i.v. each subgroup and overall, that 100-day mortality was lower in the i.v.BuCy2-MDACC patients compared to the Alt-IBMTR patients, thus comparing the two confounded treatment-center effects. The last four columns give the probability $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ for $p = 0.25$, 0.50, 0.75, and 1.00, respectively, quantifying the hypothetical treatment effects that result from assuming that 25, 50, 75, or 100% of the confounded treatment-center effect is due to center. In all three CML prognostic subgroups, even if half (50%) of the observed advantage is due to an intrinsic superiority of MDACC over the IBMTR centers, then the probability that i.v.BuCy2 has lower 100-day mortality compared to Alt preparative regimens still varies from 0.78 to 0.94, depending on prognostic subgroup. Only under the extreme assumption that 100% of the observed difference is due to MDACC center superiority over the IBMTR do the probabilities of i.v.BUCy2 treatment superiority drop to values near 0.50.

The sensitivity analyses are illustrated graphically in Figure 2. Each plot varies the assumed percentage of the confounded treatment-center effect that is due to center continuously from 0 to 100. For each CML subtype, the lowest mean center effect value of 0% corresponds to the assumption that there is no MDACC-versus-IBMTR center effect. The highest value of 100% corresponds to the assumption that, on average, all of the observed difference is due to an MDACC-versus-IBMTR advantage.

Table 2 Bayesian sensitivity analyses of center (MDACC-versus-IBMTR) effects and preparative regimen (i.v.BuCy2-versus-Alt) effects on the probabilities of 100-day mortality

	Assumed percent of estimated effect that is due to MDACC-versus-IBMTR center effect				
	0%	*25%*	*50%*	*75%*	*100%*
	Probability i.v.BuCy2 has lower 100-day mortality than Alt preparative regimens				
Prognostic subgroup					
CP	0.991	0.950	0.859	0.720	0.552
AP	>0.999	0.992	0.936	0.778	0.527
BC	0.945	0.875	0.779	0.665	0.543
Combined	0.990	0.954	0.871	0.729	0.546

Rather than first fixing the value of p, the proportion of the confounded effect that is due to superiority of MDACC, and then computing $\Pr(\pi_2 > \pi_1(p)|\text{Data})$, one may perform the computation in reverse. That is, one may first fix the posterior probability $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ at a given critical threshold and then solve for p. Table 3 provides values of p, expressed as the percentage $100p$, for the critical thresholds $\Pr(\pi_2 > \pi_1(p)|\text{Data}) = 0.80$, 0.85, 0.90, and 0.95.

More generally, one may assume a probability distribution $f(p)$ on p, the proportion of the posterior mean of $\pi_1(p)$ that may be attributed to center effect, and compute the average of $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ over $f(p)$. This formalizes

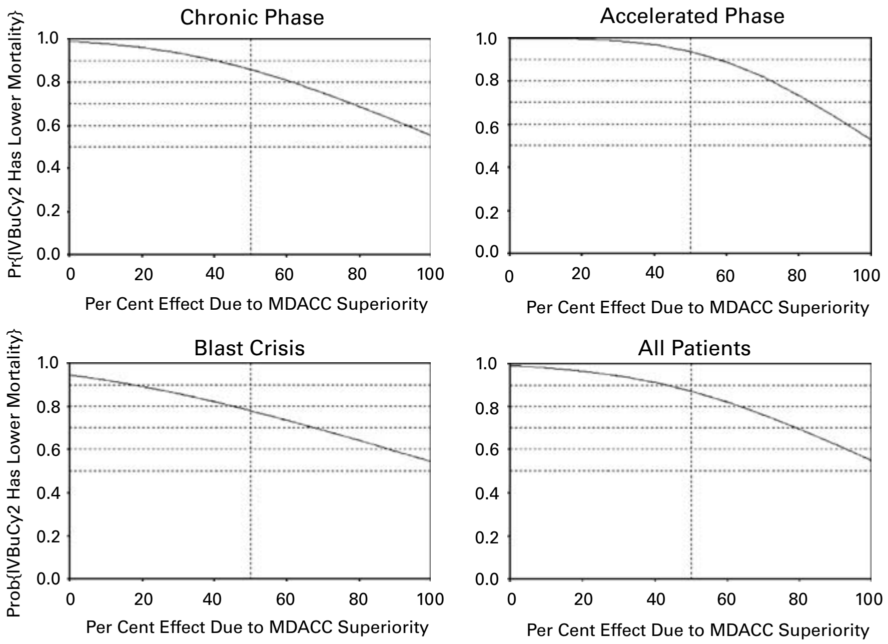

Figure 2 Bayesian sensitivity analyses. Each plot gives the posterior probability that 100-day mortality is lower with i.v.BuCy2 compared to Alt preparative regimens, as the assumed percentage of the confounded treatment-center effect that is due to center is varied between 0 and 100.

Table 3 Percent of the observed effect that is due to MDACC superiority, 100p, for given values of $\Pr(\pi_2 > \pi_1(p)|\text{Data})$, the hypothetical posterior probability that i.v.BuCy2 is superior to Alt preparative regimens

$Pr(\pi_2 > \pi_1(p)|Data)$	*Prognostic subgroup*			
	CP	*AP*	*BC*	*Combined*
0.80	61.6	72.3	45.1	63.6
0.85	52.0	65.7	32.1	54.4
0.90	40.5	57.7	17.1	43.0
0.95	24.9	46.2	0[a]	26.8

[a] $p = 0$ for the BC subgroup implies that $\Pr(\pi_2 > \pi_1(p)|\text{Data}) = 0.945$.

one's uncertainty about p, and averaging over $f(p)$ incorporates that uncertainty into the sensitivity analysis. This additional structure fits quite naturally into the Bayesian framework. From this viewpoint, each value of $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ in Table 2 may be regarded as a special case of the above in which $f(p)$ places probability 1 on a single value of p. This computation could be done for one or more hypothetical distributions on p, each reflecting a different opinion with regard to center effects. In general, the computation would require numerical integration of $\Pr(\pi_2 > \pi_1(p)|\text{Data})f(p)$ for p ranging from 0 to 1. However, one may easily do the computation in an approximate way by placing all of the probability mass of $f(p)$ on the five values of p given in Table 2 and using the values of $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ given there. For example, if one feels it is most likely that about $p = 0.50$ of the observed effect is due to MDACC superiority, but allows with some small probabilities both of the possibilities that all or none of the observed effect is due to center, then the distribution $f(0) = 0.05$, $f(0.25) = 0.10$, $f(0.50) = 0.70$, $f(0.75) = 0.10$, $f(1.00) = 0.05$ may represent this viewpoint. For this choice of $f(p)$, the average value of $\Pr(\pi_2 > \pi_1(p)|\text{Data})$ for the combined prognostic subgroups is $(0.05 \times 0.990) + (0.10 \times 0.954) + (0.70 \times 0.871) + (0.10 \times 0.729) + (0.05 \times 0.546) = 0.855$. $= 0.855$. Alternatively, the distribution $f(0) = 0.70$, $f(0.25) = 0.20$, $f(0.50) = 0.10$, $f(0.75) = f(1.00) = 0$ might reflect the viewpoint that the observed effect is most likely to be entirely due to actual treatment effect, but there is still some chance that up to half of the observed effect is due to center. For this distribution, the average is $(0.70 \times 0.990) + (0.20 \times 0.954) + (0.10 \times 0.871) + (0 \times 0.729) + (0 \times 0.546) = 0.971$.

Discussion

The statistical device of randomizing patients between two treatments provides data that may be used to construct unbiased estimates of the comparative treatment effect. While this scientific ideal is used in many clinical trials, a great deal of data result either from single-arm trials or from patients being treated with different regimens in settings where there is not complete consensus on one accepted clinical practice. Any comparison of treatments based on data *not* arising from randomized studies suffers from the confounding effects of unknown factors, such as the medical center facilities and the experience of the medical and nursing staff, selection bias, patient heterogeneity, etc. When important prognostic covariates are not available on individual patients, treatment comparisons suffer from the possibility that apparent treatment effects may, in fact, be due to differences in patient prognostic covariates. Even when such covariate data are available, there is still the concern that apparent treatment effects may be due in part to differences in clinical supportive care routines, as well as differences in training and experience of the medical and nursing staffs at the various medical centers.

The most striking feature of the early post transplant course of the 47 MDACC patients was that none died during the first 100 days post transplantation, the time period commonly considered most critical for immune recovery, and also the period during which the acute toxicity/safety of the conditioning regimen is usually assessed. In contrast, the IBMTR patients had 100-day mortality rates of 18% for those transplanted in CP, 25% in AP, and 30% in BC. Based on these data, it might appear that one may infer through standard statistical methods that i.v.BuCy2 is superior to Alt regimens with regard to 100-day mortality of CML patients receiving allotx. As explained above, however, the presence of treatment-center confounding in these data severely limits the reliability of standard statistical analyses. These problems, the absence of individual prognostic covariates for the IBMTR patients, and the possibility of supportive care effects, together motivated us to perform the Bayesian sensitivity analyses described here. We have argued that a Bayesian sensitivity analysis can provide a basis for evaluating treatment effects even in the presence of possible confounding effects.

Our Bayesian sensitivity analyses lead to the general conclusion that, if the observed differences in 100-day mortality rates are attributable to the sum of a center effect and a treatment effect, then one must conclude that i.v.BuCy2 is superior to Alt preparative regimens, that MDACC is superior to IBMTR centers, or that some combination of these two effects is the case. Since the outcome of the patients treated with i.v.BuCy2 is significantly better than what would be expected from the results available from a large number of centers participating in the IBMTR, one may be tempted to conclude that the i.v.Bu as used in the i.v.BuCy2 regimen represents a significant improvement for patient survival, at least in the first 100 days after allotx. Alternatively, if the observed difference is due to a center effect, as defined above, then one must conclude that the MD Anderson Cancer Center allotx program is superior to most other transplant programs. We are less inclined to favor the latter explanation as the most significant contributor to the observed difference in early post transplant outcome. First, post transplantation supportive care is very similar between different centers, with only minor alterations between programs. Second, the training and sharing of experience of the medical staff are similar in different programs and recruitment and exchange of staff between programs are also significant over time. The MDACC BMT program is no exception in this respect. Lastly, a number of clinical allotx studies using other myeloablative, or near-myeloablative, conditioning programs during the time period 1995–2001 have been published from the MDACC. In these publications, the 100-day mortality rate has varied from 25 to 37%.[36–43] It thus appears that the more consistent delivery of Bu that is achieved with a parenteral formulation is likely to have a significant impact on early post transplant mortality in the studied patient population. Further support for the notion that more consistent Bu delivery may impact early treatment-related mortality in CML patients undergoing allotx was recently provided by Radich *et al*.[44] These investigators reported that continuous PK monitoring and repeated Bu dose adjustments resulted in a low (3%, with 95% confidence interval 1–8%) 100-day treatment-related mortality in 131 mostly early CP patients (median time from diagnosis to BMT 5 months, with 87% transplanted within 1 year from diagnosis), who were conditioned with oral Bu and i.v.Cy in a PK-dosage-adjusted BuCy2 schedule during 1995–2000. While the inference that consistent and reliable Bu delivery is of major importance for controlling overall treatment-related mortality can be confirmed only by conducting a randomized phase III trial, until such data become available it seems appropriate when making therapeutic decisions in allotx to consider the currently available CML data in light of the results of the Bayesian sensitivity analyses reported here.

Acknowledgements

The i.v.BuCy2 studies at the MDACC were sponsored by Grants FD-R-001650-02 from the Food and Drug Administration and by CA49639 and CA16672-27 from the National Cancer Institute. Peter Thall's research was partially supported by NIH Grant R01 CA 83932. We thank two referees for their thoughtful and constructive comments on an earlier version of the manuscript.

References

1 Horowitz MM. Results of allogeneic stem cell transplantation for malignant disorders. In: Hoffman R, Benz Jr EJ, Shattil SJ, Fine B, Cohen H, Silberstein LE, McGlave P (eds). *Hematology, Basic Principles and Practice*, 3rd edn. Churchill Livingston, New York, NY, 2000; pp 1573–1587.

2 International Bone Marrow Transplant Registry. September 2000.

3 Bushhouse S, Ramsay NK, Pescovitz OH *et al*. Growth in children following irradiation for bone marrow transplantation. *Am J Ped Hematol/Oncol* 1989; **11**: 134–140.

4 Sanders JE. Bone marrow transplantation for pediatric leukemia. *Pediat Ann* 1991; **20**: 671–676.

5 Liesner RJ, Leiper AD, Hann IM *et al*. Late effects of intensive treatment for acute myeloid leukemia and myelodysplasia in childhood. *J Clin Oncol* 1994; **12**: 916–924.

6 Cohen A, van-Lint MT, Uderzo C *et al*. Growth in patients after allogeneic bone marrow transplant for hematological diseases in childhood. *Bone Marrow Transplant* 1995; **15**: 343–348.

7 Bhatia S, Ramsay NK, Steinbuch M *et al*. Malignant neoplasms following bone marrow transplantation. *Blood* 1996; **87**: 3633–3639.

8 Chou RH, Wong GB, Kramer JH *et al*. Toxicities of total-body irradiation for pediatric bone marrow transplantation. *Int J Radiat Oncol Biol Phys* 1996; **34**: 843–851.

9 Deeg HJ, Socié G, Schoch G *et al*. Malignancies after marrow transplantation for aplastic anemia and Fanconi anemia: a joint Seattle and Paris analysis of results in 700 patients. *Blood* 1996; **87**: 386–392.

10 Kony SJ, de Vathaire F, Chompret A *et al*. Radiation and genetic factors in the risk of second malignant neoplasms after a first cancer in childhood. *Lancet* 1997; **350**: 91–95.

11 Deeg HJ, Socié G. Malignancies after hematopoietic stem cell transplantation: many questions, some answers. *Blood* 1998; **91**: 1833–1844.

12 Socié G, Curtis RE, Deeg HJ *et al*. New malignant diseases after allogeneic marrow transplantation for childhood acute leukemia. *J Clin Oncol* 2000; **18**: 348–357.

13 Santos GW, Tutschka PJ, Brookmeyer R *et al*. Marrow transplantation for acute nonlymphocytic leukemia after treatment with busulfan and cyclophosphamide. *N Engl J Med* 1983; **309**: 1347–1353.

14 Lu C, Braine HG, Kaizer H *et al*. Preliminary results of high-dose busulfan and cyclophosphamide with syngeneic or autologous bone marrow rescue. *Cancer Treat Rep* 1984; **68**: 711–717.

15 Tutschka PJ, Copelan EA, Klein JP. Bone marrow transplantation for leukemia following a new busulfan and cyclophosphamide regimen. *Blood* 1987; **70**: 1382–1388.

16 Clift RA, Buckner CD, Thomas ED *et al*. Marrow transplantation for chronic myeloid leukemia: a randomized study comparing cyclophosphamide and total body irradiation with busulfan and cyclophosphamide. *Blood* 1994; **84**: 2036–2043.

17 Clift RA, Radich J, Appelbaum FR *et al*. Long-term follow-up of a randomized study comparing cyclophosphamide and total body irradiation with busulfan and cyclophosphamide for patients receiving allogeneic marrow transplants during chronic phase of chronic myeloid leukemia. *Blood* 1999; **94**: 3960–3962.

18 Grochow LB, Jones RJ, Brundrett RB *et al*. Pharmacokinetics of busulfan: correlation with veno-occlusive disease in patients undergoing bone marrow transplantation. *Cancer Chemother Pharmacol* 1989; **25**: 55–61.

19 Grochow LB. Busulfan disposition: the role of therapeutic monitoring in bone marrow transplantation induction regimens. *Semin Oncology* 1993; **20** (Suppl. 4): 18–25.

20 Hassan M, Öberg G, Ehrsson H *et al*. Pharmacokinetic and metabolic studies of high-dose busulphan in adults. *Eur J Clin Pharmacol* 1989; **36**: 525–530.

21 Hassan M, Ljungman P, Bolme P *et al*. Busulfan bioavailability. *Blood* 1994; **84**: 2144–2150.

22 Dix SP, Wingard JR, Mullins RE *et al*. Association of busulfan area under the curve with veno-occlusive disease following BMT. *Bone Marrow Transplant* 1996; **17**: 225–230.

23 Styler MJ, Crilley P, Biggs J *et al*. Hepatic dysfunction following busulfan and cyclophosphamide myeloablation: a retrospective, multicenter analysis. *Bone Marrow Transplant* 1996; **18**: 171–176.

24 Vassal G, Deroussent A, Hartmann O *et al*. Dose-dependent neurotoxicity of high-dose busulfan in children: a clinical and pharmacological study. *Cancer Res* 1990; **50**: 6203–6207.

25 Peters WP, Henner WD, Grochow LB *et al*. Clinical and pharmacologic effects of high dose single agent busulfan with autologous bone marrow support in the treatment of solid tumors. *Cancer Res* 1987; **47**: 6402–6406.

26 Vassal G. Pharmacologically-guided dose adjustment of busulfan in high-dose chemotherapy regimens: rationale and pitfalls (review). *Anticancer Res* 1994; **14**: 2363–2370.

27 Andersson BS, Madden T, Tran H *et al*. Acute safety and pharmacokinetics of intravenous busulfan when used with oral busulfan and cyclophosphamide as pretransplantation conditioning therapy: A phase I Study. *Biol Blood and Marrow Transplant* 2000; **6**: 548–554.

28 Hassan M. Busulphan. In: Grochow LB, Ames MM (eds). *A Clinical Guide to Chemotherapy Pharmacokinetics and Pharmacodynamics*. Williams and Wilkins: New York, NY, 1997; pp 189–210.

29 Benet LZ, Sheiner LB. Pharmacokinetics: The dynamics of drug absorption, distribution, and elimination. In: Goodman Gilman A, Goodman LS, Rall TW, Murad F (eds). *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 7th edn. MacMillan Publishing Co.: New York, NY, 1985; p 8.

30 Bhagwatwar HP, Phadungpojna S, Chow DS *et al*. Formulation and stability of busulfan for intravenous administration in high-dose chemotherapy. *Cancer Chemother Pharmacol* 1996; **37**: 401–408.

31 Andersson BS, Kashyap A, Gian V *et al*. Conditioning therapy with intravenous busulfan and cyclophosphamide (IV BuCy2) for hematologic malignancies prior to allogeneic stem cell transplantation: a phase II study. *Biol Blood Marrow Transplant* 2002; **8**: 145–154.

32 Andersson BS, Thall PF, Madden T *et al*. Busulfan systemic exposure relative to regimen-related toxicity and acute graft vs. host disease: defining a therapeutic window for IV BuCy2 in chronic myelogenous leukemia. *Biol Blood Marrow Transplant* 2002; **8**: 477–485.

33 Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: New York, NY, 1995.

34 Robert CP. *The Bayesian Choice*, 2nd edn. Springer Verlag: New York, NY, 2001.

35 Estey E, Thall P, Giles F *et al*. Gemtuzumab ozogamycin with or without interleukin 11 in patients 65 years of age or older with untreated acute myeloid leukemia and high-risk myelodysplastic syndrome: comparison with idarubicin + continuous-infusion high-dose cytosine arabinoside. *Blood* 2002; **99**: 4343–4349.

36 Van Besien KW, Mehra RC, Giralt SA *et al*. Allogeneic bone marrow transplantation for poor-prognosis lymphoma: response, toxicity and survival depend on disease histology. *Am J Medicine* 1996; **100**: 299–307.

37 Van Besien K, Thall P, Korbling M *et al*. Allogeneic transplantation for recurrent or refractory non-Hodgkin's lymphoma with poor prognostic features after conditioning with thiotepa, busulfan, and cyclophosphamide: experience in 44 consecutive patients. *Biol Blood Marrow Transplant* 1997; **3**: 150–156.

38 Przepiorka D, Anderlini P, Ippoliti C *et al*. Allogeneic blood stem cell transplantation in advanced hematologic cancers. *Bone Marrow Transplant* 1997; **19**: 455–460.

39 Przepiorka D, Khouri I, Thall PF *et al*. Thiotepa, busulfan and cyclophosphamide as a preparative regimen for allogeneic transplantation for advanced chronic myelogenous leukemia. *Bone Marrow Transplant* 1999; **23**: 977–981.

40 Przepiorka D, Smith TL, Folloder J *et al*. Risk factors for actue graft-versus-host disease after allogeneic blood stem cell transplantation. *Blood* 1999; **94**: 1465–1470.

41 Przepiorka D, van Besien K, Khouri I *et al*. Carmustine, etoposide, cytarabine and melphalan as a preparative regimen for allogeneic transplantation for high-risk malignant lymphoma. *Ann Oncol* 1999; **10**: 527–532.

42 Bibawi S, Abi-Said D, Fayad L *et al*. Thiotepa, busulfan, and cyclophosphamide as a preparative regimen for allogeneic transplantation for advanced myelodysplastic syndrome and acute myelogenous leukemia. *Am J Hematol* 2001; **67**: 227–233.

43 Giralt S, Thall PF, Khouri I *et al*. Melphalan and purine analog-containing preparative regimens: reduced-intensity

conditioning for patients with hematologic malignancies undergoing allogeneic progenitor cell transplantation. *Blood* 2001; **97**: 631–637.

44 Radich JP, Gooley T, Bensinger W *et al*. HLA-matched related hematopoetic cell transplantation for CML chronic phase using a targeted busulfan and cyclophosphamide preparative regimen. *Blood* 2003; **102**: 31–35.

Appendix

The Bayesian paradigm for statistical analysis is concerned with two objects, the model parameters, which we denote by θ and the observed data. Parameters may be such factors as probabilities, median survival times, or the effects of treatments or patient characteristics on a given outcome. Thus, while parameters are not observed, they characterize important aspects of the observed phenomenon. In the Bayesian framework, parameters are considered random quantities to reflect the fact that one has uncertainty about them. Consequently, a key component of the Bayesian model is a prior probability distribution on θ. The likelihood of observing the data for a given parameter θ also is a probability distribution. Bayes' Theorem formally combines one's prior with the likelihood to obtain the posterior, $f(\theta|\text{data})$, which characterizes one's uncertainty about θ, and hence about the phenomenon, after observing the data. Thus, $f(\theta|\text{data})$ is the basis for inference and decision-making in Bayesian analysis.

The methodology used here relies on two closely related probability distributions, the binomial and the beta. Consider the general setting where one observes the number of times, X, that a particular event occurs out of N independent trials, and the probability of the event in each trial is π. Then X follows a binomial probability distribution characterized by N and the parameter π and X has mean $N\pi$ and variance $N\pi(1-\pi)$. In most settings N is known, since it is simply the number of trials, but π is generally unknown. The most commonly used prior for π in such settings is the beta distribution. If π follows a beta distribution with parameters a and b, denoted $\pi \sim \text{beta}[a,b]$, then π has mean $\mu = a/(a+b)$ and variance $\mu(1-\mu)/(a+b+1)$. The sum $n = a+b$ may be interpreted as the prior sample size, so that larger n corresponds to more prior information. An equivalent, often useful way to express the beta$[a,b]$ distribution is in terms of its mean and effective sample size, μ and n, so that $\pi \sim \text{beta}[\mu n, (1-\mu)n]$. Let $Y = N-X$ denote the number of times that the event does not occur in the N trials. Once X and N have been observed, the posterior distribution of π is also beta, but with updated parameters $a+X$ and $b+Y$, denoted $\pi|X, N \sim \text{beta}[a+X, b+Y]$. The posterior mean of $\pi|X,N$ is $(a+X)/(n+N)$ and the posterior variance is $(a+X)(b+Y)/\{(n+N)^2(n+N+1)\}$. The posterior mean $(a+X)/(n+N)$ may be considered a Bayesian estimator of π, and it may be contrasted with the usual, non-Bayesian empirical mean, X/N. Some simple algebra shows that the posterior mean equals the weighted average $\mu\{n/(n+N)\} + (X/N)/\{N/(n+N)\}$ of the prior mean, $\mu = a/(a+b)$, and the empirical mean, X/N, and that the weights are proportional to the prior and actual sample sizes, n and N.

In comparing two binomial samples with probabilities π_1 and π_2 following beta distributions, we computed posterior probabilities of the form $\Pr(\pi_2 > \pi_1|\text{Data})$ using the following Splus program.

```
prob.betas <- function(a1, b1, a2, b2)

   {

 # compute Pr(pi2>pi1) where pi1~beta(a1, b1), pi2~beta(a2, b2)

integrate(intg, lower = 0, upper = 1, a1 = a1, b1 = b1, a2 = a2, b2 = b2)$integral

   }

intg <- function(p, a1, b1, a2, b2)

   {

   # the integrand of prob.betas

   pbeta(p, a1, b1) * dbeta(p, a2, b2)

   }
```