

Bayesian Sensitivity Analyses of Confounded Treatment Effects

Peter F. Thall¹ and Xuemei Wang

Department of Biostatistics and Applied Mathematics, Unit 447
University of Texas, M.D. Anderson Cancer Center
1515 Holcombe Boulevard, Houston, Texas, 77030 USA
E-mail: rex@mdanderson.org

April 23, 2004

Summary

The primary scientific goal of a randomized clinical trial of two treatments, A and B, is to compare their effects on the most important therapeutic outcome in the medical setting under study. Typically, this comparison is formulated in terms of the probabilities or hazards of the outcome with A and B, possibly transformed or covariate-adjusted. When the data arise from separate trials of A and B, however, standard statistical estimators of the A-versus-B treatment effect may be severely biased and substantively misleading due to confounding trial effects. Potential sources of systematic bias include supportive care, physician practices, covariate imbalances, and institutional effects. Although this problem is well known in the statistical community, and is the motivation for randomizing patients between treatments, it is common practice throughout the medical literature to base statistical comparisons on data from single-arm trials while ignoring trial effects. Because numerous medical decisions are based on such flawed analyses, this constitutes a serious problem, with consequences that go far beyond the scientific community. This chapter will provide illustrations of between-trial effects, and present some simple Bayesian methods for evaluating what the possible treatment effects may be in such settings.

Keywords. Bayesian statistics, bone marrow transplantation, confounding, leukemia, sensitivity analysis, survival analysis

1. Introduction

The primary scientific goal of a randomized clinical trial of two treatments, A and B, is to compare their effects on the most important therapeutic outcome in the medical setting under study. Generally, this comparison may be formulated in terms of two real-valued parameters, θ_A and θ_B , which most often are based on the probabilities or hazards of the outcome with A and B, possibly transformed or covariate-adjusted. Scientists routinely base such comparisons on the A-versus-B treatment effect, $\delta_\theta = \theta_A - \theta_B$, implicitly assuming that a typical statistical estimator computed from their data actually estimates δ_θ . While it is well established that randomization will, on average, eliminate potential sources of bias (1,2), when patients are not randomized between A and B, standard statistical estimators may become scientifically invalid and substantively misleading.

Even with randomization, in practice there are many difficulties. Patient outcome in any medical setting arises from the combined effects of treatment, patient characteristics, and other “latent” variables that are unobserved. It is well known that failure to account for patient covariates that affect outcome may produce a biased estimate of δ_θ (3). However, after accounting for known covariates through stratification, dynamic randomization (4), or regression analysis, between-patient variability still may be substantial. Because latent variables are not observed, their combined effect may appear to be attributable to treatment, thus masking an actual treatment effect. Potential sources of such systematic bias include supportive care, physician practices, or institutional effects. The randomization itself may fail in numerous ways. For example, patients who enter the trial in the hope of being treated with A may drop out if randomized to B. An adverse event more likely with A than with B may cause higher rates of drop out or early treatment termination in arm A. Physicians who favor one treatment over the other may, consciously or unconsciously, selectively withhold patients from the trial. In small to moderate sized studies, covariate imbalance may be substantial simply due to the play of chance. Still, despite the many problems imposed by practical reality, randomization remains the best scientific device for obtaining a fair treatment comparison.

The statistical problem of evaluating δ_θ becomes much more difficult when the data arise from separate trials of A and B. If A is evaluated in trial 1 and B is evaluated in trial 2, then the data from trial 1 provide an estimator not of θ_A but rather of the confounded effect, $\gamma_{A,1}$, of treatment A and trial 1. Similarly, the data from trial 2 provide an estimator of the confounded effect, $\gamma_{B,2}$, of treatment B and trial 2. A typical estimator of δ_θ based on such data actually estimates the difference between these confounded effects, $\delta = \gamma_{A,1} - \gamma_{B,2}$, rather than the treatment effect δ_θ . Of course, the distinction between δ and δ_θ is the motivation for conducting a randomized trial. There are numerous examples of between-trial effects in the statistical and medical literature. See, for example, Estey and Thall (5).

Unfortunately, it is common practice throughout the medical literature to base statistical comparisons on data from non-randomized experiments while ignoring study effects. A typical report of a single-arm trial of a treatment A may compare it to another treatment, B, based on data from one or more previous studies of B. Evidently, many scientists

believe that the use of regression methods or subset analyses to account for patient covariates when making such comparisons provides a valid estimate of δ_θ . The underlying assumptions, usually not stated explicitly, are that known covariates account for any between-trial effects (6,7), or that between-trial effects either are negligible relative to δ_θ or simply do not exist. While the statistical community regards such comparisons as fundamentally flawed, a great deal of the medical literature contains such invalid treatment comparisons. It explains, in part, why many randomized phase III trials yield negative results despite the fact that one or more preceding single-arm phase II trials indicated that the experimental treatment under study was “promising” compared to standard therapy.

This chapter has two purposes. The first is to show by example that substantial between-trial effects may persist after accounting for known covariates. Our first example will illustrate the fact that between-trial effects may be substantial even when the trials are conducted in the same institution and one accounts for covariate effects. The second purpose is to illustrate some simple Bayesian methods for assessing what the possible distribution of δ_θ may be in such settings. The basic idea underlying our approach is as follows. Let D_A and D_B denote the data from the separate trials of A and B, and denote by $f(\delta | D_A, D_B)$ the posterior of the confounded effect, $\delta = \gamma_{A,1} - \gamma_{B,2}$. Denoting the trial effects by λ_1 and λ_2 , we require the assumptions that $\gamma_{A,1} = \theta_A + \lambda_1$, the sum of effects due to treatment A and trial 1 and, similarly, that $\gamma_{B,2} = \theta_B + \lambda_2$. This implies that $\delta = (\theta_A - \theta_B) + (\lambda_1 - \lambda_2) = \delta_\theta + \delta_\lambda$, the sum of the A-versus-B treatment effect of interest and the confounding between-study effect, $\delta_\lambda = \lambda_1 - \lambda_2$. Thus, δ_θ and δ_λ are random quantities with posteriors that cannot be obtained from the available data. That is, δ_θ and δ_λ are confounded. This reveals the assumption in a typical analysis that ignores the fact that patients were not randomized between A and B, since assuming that $E(\delta) = E(\delta_\theta)$ is equivalent to assuming that $E(\lambda_1) = E(\lambda_2)$, that is, that the mean trial effects are identical. This underlies the approach of Begg and Pilote (8), and Li and Begg (9), who consider meta-analysis of combined single-arm and randomized study data. In particular, they assume that the between-trial effects for single-arm trials all have a common mean. We do not make such an assumption here.

Our first type of sensitivity analysis consists of varying hypothetical $\delta_\lambda^{(h)}$ over a reasonable set of possibilities and assessing the resulting distributions of $\delta_\theta^{(h)} = \delta - \delta_\lambda^{(h)}$. When historical data, D , from one or more clinical settings similar to the trials of A and B are available, one may vary hypothetical $\delta_\lambda^{(h)}$ over k historical between-trial effects, $\delta_{\lambda,1}, \dots, \delta_{\lambda,k}$, or possibly over a wider range of values. While this relies on the assumption that δ_λ and $\delta_{\lambda,1}, \dots, \delta_{\lambda,k}$ are stochastically similar, we do not assume that the posteriors of either δ_λ or δ_θ may be computed from historical data. We emphasize this important point by referring to the distribution $f(\delta_\theta^{(h)} | D_A, D_B, D)$ as the *hypothetical posterior* of δ_θ . When D_A and D_B are the only available data, we will take the alternative approach of fixing the variance of $\delta_\lambda^{(h)}$ and varying its mean over a specified interval.

2. Survival Analysis With Treatment-Trial Confounding

2.1. Six Leukemia Trials

Our first illustration is motivated by the desire to compare the efficacy of two combination chemotherapies for patients 65 years or older with newly diagnosed acute myelogenous leukemia (AML) or myelodysplastic syndromes (MDS). Survival time and baseline covariate data were available for patients from two separate, single-arm trials, both conducted at M.D. Anderson Cancer Center (MDACC). The first was a trial conducted in 1991-92 of the well-established combination idarubicin + cytosine arabinoside, “ara-C” (IA), and the second a subsequent trial of the newer combination gemtuzumab ozogamicin (GO). Patients in the GO trial were randomized between GO and GO plus interleuken 11 (IL-11). Additional details are given by Estey et al. (10). The IL-11 effect was negligible, and we focus on evaluating the GO-versus-IA effect, δ_{GO} . Because the AML/MDS patients were not randomized between GO and IA, any conventional statistical estimator of δ_{GO} is confounded by the between-trial effect. To deal with this, we will utilize additional data, from four other single-arm trials in AML/MDS also conducted at MDACC over the same time period, to obtain hypothetical between-trial effects. Two of the four trials were of fludarabine + IA + G-CSF (FAIG), and the other two were trials of FAIG + all-trans retinoic acid (FAIGA).

2.2. Probability Models

The patient covariates included in our survival analyses are Zubrod performance status, dichotomized as “good” = $PS \leq 2$ versus “poor” = $PS \geq 3$; whether the patient was treated in a laminar airflow room (LAR); and cytogenetic karyotype, classified into three categories: normal (the baseline group), having the very unfavorable -5/-7 abnormality, or having an abnormality other than -5/-7. We denote the linear combination of these covariates by $\beta\mathbf{Z} = \beta_0 + \beta_1 Z_1 + \dots + \beta_4 Z_4$, with β_0 the baseline hazard parameter. We denote the effect of the j^{th} treatment-trial combination compared to IA in trial 1 by δ_j , with $\delta\boldsymbol{\tau} = \delta_2 \tau_2 + \dots + \delta_6 \tau_6$ the linear term of confounded treatment-trial effects, where τ_j indicates trial $j=2, \dots, 6$. Let $S(t|\mathbf{Z}) = \text{pr}(T > t | \mathbf{Z}, \boldsymbol{\tau})$ be the survivor function, where T is survival time. We considered three possible survival time models: the Weibull, for which $\log[-\log\{S(t|\mathbf{Z})\}] = \beta\mathbf{Z} + \delta\boldsymbol{\tau} + \phi \log(t)$, the log logistic, for which $-\log[S(t|\mathbf{Z})/\{1-S(t|\mathbf{Z})\}] = \beta\mathbf{Z} + \delta\boldsymbol{\tau} + \phi \log(t)$, and the lognormal with mean $\beta\mathbf{Z} + \delta\boldsymbol{\tau}$ and constant variance. For the AML/MDS data, these models have respective maximized log likelihoods -137, -139.4 and -141.7, indicating that the Weibull gives a slightly better fit. Additionally, a plot of $\log[-\log\{S_{KM}(t)\}]$ on $\log(t)$ (not shown), where $S_{KM}(t)$ is the Kaplan-Meier estimator (11), is approximately linear, indicating that the Weibull assumption is reasonable. We thus will use this model for our sensitivity analyses.

2.3. Sensitivity Analyses of the Leukemia Data

As a preliminary analysis, we temporarily ignored the covariates and fit a simplified version of Weibull model for which $\log[-\log\{S(t|\mathbf{Z})\}] = \beta_0 + \phi \log(t)$ separately to each

trial's data. The last column of Table 1 gives the posterior mean and 95% credible interval of $\text{median}(T) = \{\exp(-\beta_0) \log(2)\}^{1/\phi}$ under this model for each trial. These estimates, which ignore trial effects and covariates, give the apparent message that, on average, the patients treated with GO in trial 2 had the worst survival.

The fit of the full Weibull model including treatment-trial effects and covariates is summarized in Table 2. The fact that $\Pr(\delta_2 > 0 | \text{data}) > 0.99$ indicates that patients given GO in trial 2 had substantially worse survival than those given IA in trial 1. Assuming that $\delta_2 = \delta_{\text{GO}} + \delta_{\lambda,2}$, the question we address in our sensitivity analysis is how much of δ_2 may have been due to the GO-versus-IA treatment effect, δ_{GO} . We will assume that each of the treatment-trial effects for trials 3,4, 5, and 6 compared to trial 1 may be decomposed into the sum of a treatment and trial effect, formally $\delta_3 = \delta_{\text{FAIG}} + \delta_{\lambda,3}$, $\delta_4 = \delta_{\text{FAIG}} + \delta_{\lambda,4}$, $\delta_5 = \delta_{\text{FAIGA}} + \delta_{\lambda,5}$, $\delta_6 = \delta_{\text{FAIGA}} + \delta_{\lambda,6}$. This allows us to exploit the fact that FAIG and FAIGA each was studied in two separate trials, so that $\delta_3 - \delta_4 = \delta_{\lambda,3} - \delta_{\lambda,4}$ and $\delta_5 - \delta_6 = \delta_{\lambda,5} - \delta_{\lambda,6}$ are covariate-adjusted between-trial effects. Since the signs of these differences are artifacts of the way the trials were indexed, $\delta_4 - \delta_3$ and $\delta_6 - \delta_5$ are also between-trial effects. The posteriors of these four between-trial effects are plotted in Figure 1. For the two FAIG trials, the posterior probability of a positive trial effect equals 0.82 or 0.18; for the two FAIGA trials these probabilities are 0.96 or 0.04. When one ignores a between-trial effect such as those in Figure 1, as is routinely done in the medical literature, this effect is added to any actual treatment effect that may exist, and the sum is incorrectly considered to be the treatment effect.

[Figure 1]

Our sensitivity analysis essentially consists of subtracting a hypothetical trial effect from the observed treatment-trial effect. For the GO-versus-IA comparison, we vary $\delta_{\lambda}^{(h)}$ among the four between-trial effects described above and evaluate the posterior of each hypothetical $\delta_{\text{GO}}^{(h)} = \delta_2 - \delta_{\lambda}^{(h)}$. The four resulting posteriors are given in Figure 2 and summarized in the first four rows of Table 3. For each of the two trial effects with negative means, it is nearly certain that GO has a higher death rate, since $\Pr(\delta_{\text{GO}}^{(h)} > 0 | \text{data}) \geq 0.99$. Even for the largest of the four trial effects, which has mean .60 and SD = .35, the odds are 7 to 3 that GO is inferior to IA. While the actual posterior distribution of δ_{GO} cannot be determined from these data, this sensitivity analysis indicates that, for each of these four hypothetical trial effects obtained from actual trials conducted at the same institution as the GO and IA trials, it is unlikely that GO provides an improvement in survival compared to IA.

[Figure 2]

A natural question is how large the hypothetical trial effect would need to be in order for $\Pr(\delta_{\text{GO}}^{(h)} > 0 | \text{data})$ to take on specified values less than 0.50, so that GO would be preferred over IA. Rows of 5 – 8 of Table 3 contain this sort of sensitivity analysis, obtained essentially by performing the previous computations in reverse. Recall that the posterior of δ_2 has mean .84 and SD .35. We equate the SD of $\delta_{\lambda}^{(h)}$ to .35, equate

$\Pr(\delta_{GO}^{(h)} > 0 \mid \text{data}) = \Pr(\delta_2 - \delta_\lambda^{(h)} > 0 \mid \text{data})$ to a value in the range from .50 to .01, and solve for $E(\delta_\lambda^{(h)})$. Table 3 shows that, in order for the probability of GO inferiority to take on a value in the range .10 to .01, one must assume a hypothetical between-trial effect having mean that is 2 to 3 times the largest mean between-trial effect of .60 actually seen.

3. Analyzing Confounded Count Data

3.1. Data and Problem Definition

The data summarized in Table 2 arose from two sources. The first was a single-arm clinical trial of intravenous Busulfan and Cytosin (IVBuCy) as a conditioning regimen for 47 allogeneic blood or marrow transplantation (allotx) patients with chronic myelogenous leukemia (CML). This trial was conducted at M.D. Anderson Cancer Center (MDACC) from July 1996 to October 1999. Additional details are provided by Andersson et al. (12). The second data source was the International Bone Marrow Transplantation Registry (IBMTR), which provided cross-tabulated counts of prognostic category and the indicator of 100-day mortality for 1765 CML patients who received allotx with alternative preparative regimens (Alt), primarily Cytosin with total body irradiation or oral Busulfan and Cyclophosphamide. The three CML disease stages are chronic phase (CP), accelerated phase (AP), and blast crisis (BC). Table 4, which summarizes the data, illustrates the well-known fact that the probability of short-term survival in CML patients undergoing allotx decreases ordinally with disease stage, with CP the best and BC the worst prognostic stage. While other covariates were available for the MDACC patients, after accounting for disease stage none were of any additional value for predicting the probability of 100-day survival. No patient prognostic covariates other than disease stage were available for the IBMTR patients. The data that we will utilize for comparison of the two conditioning regimens, IVCy and Alt, consist of 100-day mortality counts given in Table 4.

The main scientific difficulty with these data arises from the fact that patients were not randomized between IVCy and Alt. Instead, all 47 IVCy patients were transplanted at MDACC while all 1765 Alt patients were transplanted at IBMTR medical centers, so the two preparative regimens are confounded with the centers. We will denote these two confounded treatment-center groups as IVCy-MDACC for IVCy at MDACC and Alt-IBMTR for Alt regimens at the IBMTR centers. Within each row of Table 4, each comparison reflects the difference between these confounded treatment-center effects, rather than the IVCy-versus-Alt treatment effect. Thus, these preliminary comparisons are potentially misleading in that they ignore treatment-center confounding.

A visual inspection of the event rates shows the obvious and important fact that none of the IVCy-MDACC patients died within the first 100 days post transplant, whereas the 100-day mortality rates for the Alt-IBMTR patients varied from 18% to 30%, depending on CML stage. In addition to treatment-center confounding, two other interesting aspects of these data are that all of the IVCy-MDACC counts are 0, and the IVCy-MDACC sample size of 47 is much smaller than the 1745 Alt-IBMTR sample size. In summary, the problem is to compare treatment effects between two sets of binomial samples

accounting for prognostic subgroup, in the presence of treatment-center confounding, based on disproportionate sample sizes where all counts in the smaller sample are 0.

3.2. Probability Model

The following sensitivity analysis is similar to that given by Thall, Andersson and Champlin (13). Index the CML stage subgroups by $j=CP, AP, BC$, and denote the 100-day mortality probabilities in subgroup j by $\gamma_{1,j}$ for IVBuCy-MDACC patients and $\gamma_{2,j}$ for Alt-IBMTR patients. We assume *a priori* that each $\gamma_{i,j} \sim \text{iid beta}(.5, .5)$, so that each prior contains as much information as knowing whether one patient died within 100 days. Denoting the data by $X_{i,j} = \# \text{ deaths}$ and $N_{i,j} = \# \text{ patients}$, the posterior of $\gamma_{i,j}$ is $\text{beta}(.5+X_{i,j}, .5+N_{i,j}-X_{i,j})$, which has posterior mean $\mu_{i,j} = (X_{i,j}+.5) / (N_{i,j}+1)$ and variance $\sigma_{i,j}^2 = \mu_{i,j}(1-\mu_{i,j})/(N_{i,j}+2)$. Observing that none of the 17 IVBuCy-MDACC patients in CP died within 100 days (Table 4) gives a $\text{beta}(.5,17.5)$ posterior for $\gamma_{1,CP}$, which has mean .028 (sd = .038). Observing that 242/1344 Alt-IBMTR CP patients died within 100 days gives a $\text{beta}(242.5,1102.5)$ posterior for $\gamma_{2,CP}$, which has mean 0.180 (sd = .010), reflecting both the much higher observed 100-day mortality rate and the much larger sample size. The posteriors for the other prognostic groups are computed analogously. These posteriors are graphed in Figure 3, which illustrates the fact that, since 0/47 IVBuCy-MDACC patients died within 100 days, within each prognostic subgroup and overall they had a much smaller posterior probability of 100-day mortality than the Alt-IBMTR patients.

[Figure 3]

A Bayesian statistic for comparing $\gamma_{2,j}$ and $\gamma_{1,j}$ is $\Pr(\gamma_{2,j} > \gamma_{1,j} | \text{data}_j)$, the posterior probability that the 100-day mortality rate in Alt-IBMTR patients is higher than in IVBuCy-MDACC patients in subgroup j . This probability is .50 if the posteriors of $\gamma_{1,j}$ and $\gamma_{2,j}$ are identical, and values greater (less) than 0.50 correspond to a lower (higher) 100-day death rate in the IVBuCy-MDACC patients. A single overall probability may be obtained by averaging the probabilities of the individual subgroups. Weighting proportional to sample size (Table 4), the sample proportions are $w_{CP} = .75$, $w_{AP} = .20$, and $w_{BC} = .05$, so the average is $\sum_{j=CP,AP,BC} w_j \Pr(\gamma_{2,j} > \gamma_{1,j} | \text{data}_j)$. The values of $\Pr(\gamma_{2,j} > \gamma_{1,j} | \text{data}_j)$ for $j=CP, AP, BC$ and the weighted average are given in the last column of Table 4. These indicate that, assuming uninformative priors, it is virtually certain *a posteriori* that IVBuCy-MDACC had a lower 100-day mortality probability than Alt-IBMTR for the CP patients, for the AP patients, on average across the three prognostic groups, and the odds are about 17 to 1 in favor of IVBuCy-MDACC for the BC patients. The issue now is what may be said about the IVBuCy-Alt treatment effect based on these data.

3.3. Sensitivity Analyses of the Transplant Data

First consider a single disease subtype. For convenience, temporarily suppress the index j , and consider the confounded effect, $\delta = \gamma_2 - \gamma_1$, of Alt-IBMTR versus IVBuCy-MDACC on 100-day mortality. Let p be the hypothetical proportion of δ accounted for by center, for $0 \leq p \leq 1$, so that the remaining $1-p$ is due to treatment, and denote by $\theta_1(p)$ the

hypothetical 100-day mortality probability of an IVBuCy patient if center effect could be completely removed. Since $\gamma_1 | \text{data} \sim \text{beta}\{\mu_1 N_1, (1-\mu_1)N_1\}$ with $\mu_1 = (\frac{1}{2} + X_1)/(1 + N_1)$, the posterior effective sample size of γ_1 is $M_1 = N_1 + 1$. Define the weighted average $\mu_1(p) = (1-p)\mu_1 + p\mu_2$ of the posterior means. We assume that $\theta_1(p) | \text{data} \sim \text{beta}[\mu_1(p)M_1, \{1-\mu_1(p)\}M_1]$. This says that, for each value of p , the hypothetical posterior 100-day mortality probability of an IVBuCy patient, with center effect removed, follows a beta distribution with mean $\mu_1(p)$ and effective sample size the same as that of γ_1 . The decomposition $\delta = \{\gamma_2 - \theta_1(p)\} + \{\theta_1(p) - \gamma_1\} = \delta_\theta^{(h)}(p) + \delta_\lambda^{(h)}(p)$ formalizes the assumption that $\{\text{confounded effect}\} = \{\text{Alt-versus-IVBuCy treatment effect}\} + \{\text{IBMTR-versus-MDACC center effect}\}$. Thus, $\delta_\theta^{(h)}(p) = \gamma_2 - \theta_1(p)$ is the hypothetical IVBuCy-Alt treatment effect under the assumption that $p100\%$ of δ is due to treatment. At one extreme, if $p=0$ then there is no center effect, $\theta_1(p) = \theta_1(0)$ has the same distribution as γ_1 , $\Pr(\gamma_2 > \theta_1(p) | \text{data}) = \Pr(\gamma_2 > \gamma_1 | \text{data})$, and $\delta = \delta_\theta$. If $p = .5$, then $\theta_1(p) = \theta_1(.5)$ has mean $.5\mu_1 + .5\mu_2$, and on average half of the observed difference is due to treatment and half to center. If $p=1$, then $\theta_1(p) = \theta_1(1)$ has the same mean as γ_2 , all of the observed effect is due to center, there is no treatment effect, and $\Pr(\gamma_2 > \theta_1(p) | \text{data})$ is approximately .. This probability is not exactly .5 because γ_2 and $\theta_1(p)$ have different variances. Re-introducing j , our sensitivity analysis will consist of evaluating $\Pr(\delta_{\theta,j}^{(h)}(p) > 0 | \text{data}) = \Pr(\gamma_{2,j} > \theta_{1,j}(p) | \text{data})$ as a p is varied from 0 to 1 for each $j = \text{CP, AP, BC}$, and also for the weighted average of the three subgroups.

The sensitivity analyses are summarized in Table 5. The first column of the table, labeled “0%”, gives the probability $\Pr(\gamma_2 > \gamma_1 | \text{data})$, within each subgroup and overall, that 100-day mortality was lower in the IVBuCy-MDACC patients compared to the Alt-IBMTR patients, thus comparing the two confounded treatment-center effects. The last four columns give $\Pr(\gamma_2 > \theta_1(p) | \text{data})$ for $p = .25, .50, .75$, 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 50% of the observed advantage is due to an intrinsic superiority of MDACC over the IBMTR centers, then the probability that IVBuCy has lower 100-day mortality compared to alternative preparative regimens still varies from .78 to .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 IVBuCy treatment superiority drop to values near .50.

More generally, one may assume a probability distribution $f(p)$ on p and compute the average $\int \Pr(\gamma_2 > \theta_1(p) | \text{data}) f(p) dp$. This formally incorporates one’s uncertainty about p into the sensitivity analysis. Each value of $\Pr(\gamma_2 > \theta_1(p) | \text{data})$ in Table 5 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 may be carried out for each of several hypothetical distributions on p , reflecting different opinions regarding center effects, and the sensitivity analysis would then be with regard to the different values of f . To avoid numerical integration of $\Pr(\gamma_2 > \theta_1(p) | \text{data}) f(p)$ over p , one may approximate the integral by placing all of the probability mass of $f(p)$ on a few values of p . We did this using the five values $p = .05, .25, .50, .75, .95$. For example, if one feels it is most likely that about $p=.50$ 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 one might assume $f(0) = .05$, $f(.25) = .10$, $f(.50) = .70$, $f(.75) = .10$, $f(1.00) = .05$. For this choice of $f(p)$, the average value of $\Pr(\gamma_2 > \theta_1(p)|\text{data})$ for the combined prognostic subgroups equals $.05*.990 + .10*.954 + .70*.871 + .10*.729 + .05*.546 = .855$. Alternatively, the distribution $f(0) = .70$, $f(.25) = .20$, $f(.50) = .10$, $f(.75) = f(1.00) = 0$ reflects 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 of $\Pr(\gamma_2 > \theta_1(p)|\text{data})$ is .971. With this more general approach, the sensitivity analysis consists of evaluating $\int \Pr(\gamma_2 > \theta_1(p)|\text{data})f(p)dp$ as a function of f , as f is varied over a reasonable range of distributions.

An alternative approach that is very similar to assuming that $\delta = \delta_\theta^{(h)}(p) + \delta_\lambda^{(h)}(p)$ and varying p is to assume that a given number of IBMTR deaths are due to center and that the rest are due to an actual Alt-versus-IVBuCy treatment effect, and vary this assumed number. Thus, $\Pr(\gamma_2 > \theta_1(p)|\text{data})$ is replaced by $\Pr(\gamma_2 > \gamma_1 | \text{data}^{(h)})$. For example, in the 1344 Alt-IBMTR CP patients, as the number of deaths assumed to be due to center is varied from 0 to the observed 242, one obtains results similar to those given in Table 5. If one assumes that, respectively, 60 (25%), 121 (50%), or 182 (75%) of the 242 deaths are due to IBMTR-MDACC center differences, the corresponding values of $\Pr(\gamma_2 > \gamma_1 | \text{data}^{(h)})$ are .975, .928, and .790. This type of analysis is conceptually similar to, but not the same as, the sensitivity analysis of hypothesis tests based on attributed effects in observational data described by Rosenbaum (14).

4. Discussion

The type of Bayesian sensitivity analyses described here are no substitute for conducting comparative clinical trials correctly in the first place. However, because physicians and scientists conduct many single-arm clinical trials and other experiments, biostatisticians are routinely confronted with the problem of comparing treatments based on data from non-randomized studies. One solution is simply to refuse to use such data for treatment comparison and insist that a randomized trial be conducted. This leaves the problem to be solved by non-statisticians, who likely will analyze the data using statistical methods that implicitly assume the data arose from a randomized experiment. Once such results are published, the consequence is that subsequent medical practice is not unlikely to be based on fallacious conclusions. That is, ignoring the problem will not make it go away. Our proposed approach is simply to assess, as honestly as possible, what the distribution of a treatment effect may be under reasonable assumptions about confounding effects.

Acknowledgements

Peter Thall's research was partially supported by NIH grant R01 CA 83932.

References

1. Fisher, RA. Design of Experiments. Edinburgh: Oliver and Boyd, 1935.

2. Meier P. Statistics and medical experimentation. *Biometrics* 1975; 31:511-529.
3. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; 71:431-444.
4. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial (Corr: V32 p954-955)', *Biometrics* 1975; 31:103-115.
5. Estey EH, Thall, PF. New designs for phase 2 clinical trials. *Blood* 2003; 102: 442-448.
6. Gehan EA and Freireich EJ. Non-randomized controls in cancer clinical trials. *New England Journal of Medicine* 1974; 290:198-203.
7. Gehan, EA. The evaluation of therapies: Historical control studies. *Statistics in Medicine* 1984; 3:315-324.
8. Begg CB and Pilote L. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991; 47:899-906.
9. Li Z, Begg CB. Random effects models for combining results from controlled and uncontrolled studies in meta-analysis. *Journal of the American Statistical Association* 1994; 89:1523-1527.
10. Estey EH, Thall PF, Giles F, Wang, XM, Cortes JE, Beran M, Pierce SA, Thomas DA, Kantarjian HM. 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.
11. Kaplan EL, Meier P. Nonparametric estimator from incomplete observations. *Journal of the American Statistical Association* 1958; 53:457-481.
12. Andersson BS, Thall PF, Madden T, Couriel D, Wang Z, Tran HT, Anderlini P, deLima M, Gajewski J, Champlin RE. 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. *Biology of Blood and Marrow Transplantation* 2002; 8:477-485.
13. Thall PF, Andersson BS, Champlin RE. Comparison of 100-day mortality rates associated with IV busulfan and cyclophosphamide versus other preparative regimens in allogeneic bone marrow transplantation for chronic myelogenous

leukaemia: Bayesian sensitivity analyses of confounded treatment and center effects. *Bone Marrow Transplantation* 2004; vv:pp-pp.

14. Rosenbaum, PR. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* 2001; 88:219-231.

Table 1. Treatment and trial effects for the six AML/MDS trials. The four treatment combinations are IA = idarubicin + ara-C, GO = gemtuzumab ozogamicin \pm IL-11, FAIG = fludarabine + IA + G-CSF, and FAIGA = FAIG + ATRA. The effect of treatment combination C versus IA is denoted by δ_C , and the effect of trial k versus trial 1 is denoted by $\delta_{\lambda,k}$.

Trial	Treatment	Estimable Effects	Assumed Treatment + Trial Effects	# Deaths / # Patients	Posterior Mean (95% Credible Interval) of Median Survival
1	IA	—	—	16/31	47 (20-105)
2	GO	δ_2	$\delta_{GO} + \delta_{\lambda,2}$	29/51	12 (7-93)
3	FAIG	δ_3	$\delta_{FAIG} + \delta_{\lambda,3}$	34/36	14 (7-24)
4	FAIG	δ_4	$\delta_{FAIG} + \delta_{\lambda,4}$	18/22	30 (13-63)
5	FAIGA	δ_5	$\delta_{FAIGA} + \delta_{\lambda,5}$	33/44	37 (20-64)
6	FAIGA	δ_6	$\delta_{FAIGA} + \delta_{\lambda,6}$	12/17	53 (18-128)

Table 2. Fitted Bayesian Weibull regression model for survival time. Aside from ϕ , a positive parameter value is associated with shorter survival.

Posterior Distribution		
Variable	Mean_{SD}	95% Credible Interval
Covariates		
Intercept	0.62 _{0.29}	(-1.21, -0.09)
Performance status = 3 or 4	0.67 _{0.24}	(0.20, 1.14)
Treatment in laminar airflow room	-1.04 _{0.21}	(-1.45, -0.61)
-5/-7 cytogenetic abnormality	1.23 _{0.24}	(0.77, 1.71)
Other cytogenetic abnormalities	0.63 _{0.24}	(0.18, 1.09)
Treatment-Trial Effects, versus IA in Trial 1		
Mylotarg in trial 2 (δ_2)	0.84 _{0.33}	(0.20, 1.48)
FAIG in trial 3 (δ_3)	0.74 _{0.35}	(0.10, 1.45)
FAIG in trial 4 (δ_4)	0.47 _{0.38}	(-0.27, 1.23)
FAIG + ATRA in trial 5 (δ_5)	0.39 _{0.34}	(-0.25, 1.08)
FAIG + ATRA in trial 6 (δ_6)	-0.21 _{0.40}	(-0.96, 0.61)
Shape parameter (ϕ)	0.71 _{0.05}	(0.63, 0.81)

Table 3. Assumed hypothetical trial effect $\delta_\lambda^{(h)}$, the corresponding hypothetical effect $\delta_{GO}^{(h)} = \delta_2 - \delta_\lambda^{(h)}$ of GO versus IA, and the hypothetical posterior probability $\Pr(\delta_{GO}^{(h)} > 0 | \text{data})$ that GO is associated with worse survival compared to IA.

Assumed trial effect mean _{SD}	Corresponding GO effect mean _{SD}	Posterior probability that GO is inferior to IA
-0.27 _{0.29}	1.11 _{0.44}	0.99
0.27 _{0.29}	0.56 _{0.44}	0.9
-0.60 _{0.35}	1.44 _{0.50}	>0.99
0.60 _{0.35}	0.23 _{0.47}	0.69
0.84 _{0.35}	0 _{0.48}	0.50
1.45 _{0.35}	-0.61 _{0.48}	0.10
1.62 _{0.35}	-0.78 _{0.48}	0.05
1.94 _{0.35}	-1.10 _{0.48}	0.01

Table 4. 100-day survival of 1,812 CML patients for each confounded preparative regimen-medical center combination.

Prognostic Subgroup	# Deaths within 100 days / # Patients (%)		Posterior probability that IVBuCy-MDACC has lower 100-day mortality than Alt-IBMTR
	IVBuCy at MDACC	Alternative Preparative Regimens at IBMTR Centers	
Chronic Phase	0 / 17 (0)	242 / 1344 (18)	0.991
Accel. Phase	0 / 25 (0)	84 / 335 (25)	> 0.999
Blast Crisis	0 / 5 (0)	26 / 86 (30)	0.945
Overall	0 / 47 (0)	352 / 1765 (20)	0.990

Table 5. Bayesian sensitivity analyses of center (MDACC-versus-IBMTR) effects and preparative regimen (IVBuCy-versus-Alt) effects on the probabilities of 100-day mortality. Each entry is the probability $\Pr(\gamma_2 > \theta_1(p) | \text{data})$ that IVBuCy has lower 100-day mortality than Alternative preparative regimens.

Prognostic Subgroup	Assumed hypothetical proportion p of the confounded effect that is due to MDACC-versus-IBMTR center effect				
	0	.25	.50	.75	1.00
Chronic Phase	.991	.950	.859	.720	.552
Accel. Phase	> .999	.992	.936	.778	.527
Blast Crisis	.945	.875	.779	.665	.543
Overall	.990	.954	.871	.729	.546

Figure legends

Figure 1. The posterior trial effect distributions derived from the two FAIG trials (3 and 4) and the two FAIG+ATRA trials (5 and 6).

Figure 2. Posterior distributions of the hypothetical GO-versus-IA effects for the assumed hypothetical trial effects in Table 3. The hypothetical probability that GO is associated with worse survival than IA is $p = \Pr(\delta_{GO}^{(h)} > 0 | \text{data})$.

Figure 3. Posterior distributions of the IVBUCy-MDACC and Alt-IBMTR 100-day mortality probabilities, within each prognostic subgroup and overall.