Biostatistics (2003), **4**, 4, *pp.* 495–512
Printed in Great Britain

Incorporating prior beliefs about selection bias into the analysis of randomized trials with missing outcomes

DANIEL O. SCHARFSTEIN[†]
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
dscharf@jhsph.edu

MICHAEL J. DANIELS
Department of Statistics, University of Florida, Gainesville, FL 32611, USA

JAMES M. ROBINS
Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

SUMMARY

In randomized studies with missing outcomes, non-identifiable assumptions are required to hold for valid data analysis. As a result, statisticians have been advocating the use of sensitivity analysis to evaluate the effect of varying asssumptions on study conclusions. While this approach may be useful in assessing the sensitivity of treatment comparisons to missing data assumptions, it may be dissatisfying to some researchers/decision makers because a single summary is not provided. In this paper, we present a fully Bayesian methodology that allows the investigator to draw a 'single' conclusion by formally incorporating prior beliefs about non-identifiable, yet interpretable, selection bias parameters. Our Bayesian model provides robustness to prior specification of the distributional form of the continuous outcomes.

Keywords: Dirichlet process prior; Identifiability; MCHC; Non-parametric Bayes; Selection model; Sensitivity analysis.

1. INTRODUCTION

In randomized studies with missing outcomes, it is well known that non-identifiable assumptions (e.g. missing at random; Rubin, 1976) are required to hold for valid data analysis. The degree to which these untestable assumptions are believed can have a substantial impact on study conclusions. With this in mind, statisticians have been advocating the use of sensitivity analysis to evaluate the effect of varying assumptions on study conclusions. For example, Rotnitzky *et al.* (1998, 2001), Scharfstein *et al.* (1999), Robins *et al.* (2000) adopted a selection modeling approach; while Rubin (1977), Little (1994) and Daniels and Hogan (2000) used a pattern-mixture formulation. These approaches rely heavily on expert opinions about plausible ranges for non-identifiable, yet interpretable, sensitivity analysis parameters.

While the above methodological developments are useful in assessing the sensitivity of treatment comparisons to missing data assumptions, it may be dissatisfying to some researchers/decision makers

[†]To whom correspondence should be addressed

because a single summary is not provided. A fully Bayesian analysis allows the investigator to draw a 'single' conclusion by formally incorporating prior beliefs about model parameters. For categorical outcomes, Robins *et al.* (1999) and Raab and Donnelly (1999) developed fully Bayesian selection modeling approaches, while Forster and Smith (1998) developed a pattern-mixture approach. For continuous outcomes, Lee and Berger (2001), building on the work of Bayarri and Degroot (1987) and Bayarri and Berger (1998), developed a semiparametric Bayesian selection modeling approach, which places strong distributional assumptions on the outcome and weak distributional assumptions on the selection mechanism. In this paper, we consider the continuous outcome setting but take an opposite tack from Lee and Berger (2001). That is, we place strong prior restrictions on the selection mechanism, but relax the distributional restrictions on the outcome. Our tack is motivated by the fact that, in the clinical trial setting, investigators may have firmer beliefs about the selection mechanism as opposed to the distributional form of the outcome. The flexibility we seek makes the problem challenging. As a result, we restrict ourselves to the setting in which additional covariate information is ignored. By closely examining this scenario, we will gain insight into the more difficult and realistic setting, in which covariate information is utilized. This latter setting will be addressed in a sequel.

The paper is organized as follows. In Section 2, we describe an AIDS clinical trial which will provide context for the methods discussed throughout. In Section 3, we formalize the data structure of the AIDS study. In Section 4, we review the frequentist, non-parametric sensitivity analysis approach of Rotnitzky and colleagues. This review provides a backdrop for our flexible Bayesian approach, developed in Section 5. In Section 6, we analyze the AIDS data from both the frequentist and Bayesian perspective and compare results. Section 7 is devoted to a discussion.

2. ACTG 175

ACTG 175 was a randomized, double-blind trial designed to evaluate nucleoside monotherapy versus combination therapy in HIV-infected individuals with CD4 counts between 200 and 500 mm^{-3}. 2467 subjects were randomized to one of four treatment arms: 619 to AZT (600 mg a day) alone, 613 to AZT (600 mg a day) + ddI (400 mg a day), 615 to AZT (600 mg a day) + ddC (2.25 mg a day), and 620 to ddI (400 mg a day) alone (Hammer *et al.*, 1996). CD4 counts were scheduled to be collected at baseline, week 8, and then every 12 weeks thereafter. Additional baseline characteristics were also collected. In the interest of space, we focus attention on the AZT+ddI and ddI treatment arms. Also, we ignore all recorded information except the CD4 count to be measured at week 56.

One goal of the investigators was to compare the treatment-specific distributions of CD4 cell count at week 56 had all subjects remained on their assigned treatment through that week. Thus, it is useful to define a completer as a subject who stays on therapy and is measured at week 56; otherwise, we define the subject as a drop-out. In this paper, we do not distinguish between the multiple causes of drop-out. The percentage of drop-outs in the AZT+ddI and ddI arms is 33.6% and 26.5%, respectively. To address the above objective, a completers-only analysis is usually performed. The mean CD4 count at week 56 for completers (standard error) is 384.96 (8.53) and 359.59 (7.67) in the AZT+ddI and ddI arms, respectively. The difference in means is 25.36 and the associated 95% confidence interval is (2.87, 47.85); a test of the null hypothesis of no treatment difference has an associated p-value of 0.027, taken to be evidence of the superiority of AZT+ddI over ddI. The above estimates of the means, under full completion, are only valid if the completers and drop-outs are similar on measured (ignored) and unmeasured characteristics (i.e. missing at random). This latter, non-identifiable assumption is unlikely to hold, as it is well known from other studies that drop-outs tend to be very different than completers. Our goal is to present two alternative and complementary analysis strategies for the ACTG 175 data. The first approach is frequentist, while the second is Bayesian.

3. Data structure and notation

We focus on an individual treatment group. We let Y denote the CD4 count that would be observed at week 56 under full compliance with assigned therapy. Let R be the completion indicator, so that $R = 1$ if the subject is a completer and $R = 0$ if he is a drop-out. Thus, Y is observed when $R = 1$ and missing when $R = 0$. Ignoring all other recorded information, we think of $C = (R, Y)$ and $O = (R, Y : R = 1)$ as the complete and observed data for an individual, respectively. We assume that $\{C_i = (R_i, Y_i) : i = 1, \ldots, n\}$ and $\{O_i = (R_i, Y_i : R_i = 1) : i = 1, \ldots, n\} \equiv \boldsymbol{O}$ are sets of n independent and identically distributed (iid) copies of C and O, respectively.

Let f be the probability density function (pdf) of Y, f_1 be the conditional pdf of Y among completers, and f_0 be the conditional pdf of Y among drop-outs. Let F, F_1, and F_0 be the corresponding cumulative distribution functions (cdf). Let $p = P[R = 1]$. With this notation, note that the observed data law, G_O, is characterized by p and F_1.

The goal is to use $\boldsymbol{O}$ to draw inference about a functional of F, $\mu(F)$. The main functional of interest is $E[Y]$ ($\mu(F) = \int y \, \mathrm{d}F(y)$).

4. Frequentist inference

Here, we review the main results of the non-parametric, sensitivity analysis methodology of Rotnitzky and colleagues. Our exposition is intended to provide background and motivation for the Bayesian development in Section 5.

4.1 *Identifiability, non-parametric models, and sensitivity analysis*

It is well known that the law G_O of the observed data O is not sufficient to identify the distribution of Y (f) or any of its functionals. This is because (1) f can be expressed as a mixture of the conditional distributions of Y for completers (f_1) and drop-outs (f_0), weighted by the probability of completion (p) and (2) the distribution of Y among drop-outs (f_0) is not identified from G_O.

One way to identify the distribution of Y is to place just enough restrictions on the complete data laws to identify f, without restricting the laws of the observed data. Towards this end, Rotnitzky and colleagues assume a relationship between f_0 and f_1. In particular, they specify a function q and postulate that

$$f_0(y) = f_1(y) \frac{\exp(q(y))}{\int_{-\infty}^{\infty} exp(q(s)) \, \mathrm{d}F_1(s)} \quad \forall \; y. \tag{1}$$

It is important to recognize that implicit in the above relationship is the non-identifiable assumption that the support of the distribution of Y among completers is the same as that for drop-outs.

For any G_O and each q, (1) identifies a unique law of the complete data C, G_C, which marginalizes to G_O. That is, (1) generates a one-to-one q-dependent mapping between G_O and G_C, where G_C satisfies (1). Since this holds for any G_O, positing q in (1) places no restrictions on the laws of the observed data. Thus, for each q, assumption (1) is a non-parametric model for the observed data. Also, the function q is not identified from G_O since the same observed data likelihood is generated for all q. Under (1) with specified q, the marginal cdf of Y is identified via the following formula:

$$F(y) = F_1(y)p + \frac{\int_{-\infty}^{y} \exp\left(q\left(s\right)\right) \mathrm{d}F_1(s)}{\int_{-\infty}^{\infty} \exp\left(q\left(s\right)\right) \mathrm{d}F_1(s)}(1 - p) \equiv \Phi_q(p, F_1)(y) \quad \forall \; y. \tag{2}$$

The restriction that (1) places on the laws of the complete data is identically equivalent to the following logistic regression restriction:

$$\text{logit } P[R = 0|Y] = \eta + q(Y) \tag{3}$$

where η is an unknown scalar parameter. Thus, there exists a one-to-one q-dependent mapping between G_O and G_C, where G_C satisfies (3).

In the frequentist paradigm, there is thought to be one true function q which generated the complete data. However, the observed data contain no evidence about this function q. Within the context of this paradigm, Rotnitzky and colleagues recommended repeating the analysis over a range of q judged plausible by field experts. The following section discusses the two ways in which q can be interpreted. This may help in the elicitation of reasonable ranges from the experts. Finally, it is important to note that q would be partially or wholly identified if additional modeling assumptions were imposed. For example, suppose that it is assumed that the marginal distribution of Y was symmetric. Then, when q is a constant function, f_1 must be symmetric. If the empirical distribution of Y among completers is skewed, then we have evidence q is non-constant. In Section 6, we show that assuming $\log(Y)$ follows a normal distribution is sufficient to identify the magnitude of selection bias. Our belief is that it is rare that one would have such strong *a priori* knowledge about the distribution of Y that it should be used to identify q.

4.2 *Interpretation and parametrization of q*

From (1), we see that q indicates how the distribution of Y among completers relates to the distribution of Y among drop-outs. Equation (3) tells us that q quantifies the influence of the outcome Y on the odds that subjects drop out. From this latter interpretation, we refer to q as a selection bias function.

Positing that $q = 0$ is equivalent to missing at random, which in this setting is the same as the missing completely at random assumption (Rubin, 1976). This equivalence follows since the selection model (3) does not then depend on the observed data. Using the pattern mixture representation (1), $q = 0$ says that the distribution of Y among drop-outs is the same as that of completers. From the selection bias representation, $q = 0$ says that Y has no influence on a subject's completion probability. When q is non-constant, the outcome Y is said to be missing not at random.

There are obviously an infinite number of choices for q. In conducting a sensitivity analysis, it is useful to restrict attention to a simple class of selection bias functions, which include missing at random. In addition, subject-matter experts need a clear and meaningful parametrization of the selection bias function in order to encode their beliefs. The dimension of the parametrization needs to be relatively low, because otherwise researchers may be hard pressed to encode their beliefs in high dimensions. While some analysts may feel uncomfortable focusing on a particular parametrization, it is important to recognize that the aim is not to find the 'truth' (since it is unknowable without random follow-up sampling), but to report an analysis which reasonably reflects an expert's beliefs about selection bias.

In our data analysis, we consider the class of functions $\mathcal{Q} = \{\alpha \log(Y) : \alpha \in R\}$, indexed by a selection bias parameter α (note that we could have considered alternative parametrizations to reflect additional non-linearities considered plausible by experts, e.g. piecewise linear). In $\mathcal{Q}$, $\alpha = 0$ is equivalent to missing at random and $\alpha \neq 0$ is equivalent to missing not at random. For given α, we denote the mapping from the observed data distribution to the complete data distribution by Φ_α. From (3), the parameter α is interpreted as the log odds ratio of drop-out between subjects who differ by one unit of $\log(Y)$. So, $\alpha > 0$ (< 0) indicates that subjects with higher (lower) CD4 counts under full compliance are more likely to drop-out. For example, $\alpha = 0.5(-0.5)$ implies that a k-fold ($k > 1$) increase in CD4 count at week 56 leads to a $k^{0.5}$-fold increase (decrease) in the odds of drop-out. From (1), we see that $\alpha > 0$ (< 0) indicates that the distribution of Y among drop-outs is more (less) heavily weighted towards high values of Y than the

distribution of Y among completers. To make this more concrete, in Figure 1 we present the treatment-specific imputed distributions for Y among drop-outs for various values of α. These distributions were found by plugging the empirical histogram of Y among completers into the right-hand side of (1). We see that when $\alpha = 0$, the imputed distribution is, as expected, exactly equal to the distribution of Y among completers. When α is negative (positive), we see that the imputed distribution is more heavily weighted towards low (high) CD4 counts, indicating that sicker (healthier) subjects are the ones who are dropping out. The degree of weighting increases as α becomes more extreme. Based on previous studies, it has been observed that sicker subjects tend to drop out (i.e. $\alpha < 0$).

4.3 *Estimation and large sample theory*

To estimate F, under restriction (1, 3), we use non-parametric maximum likelihood. We estimate $F(y)$ by $F_n(y) = \Phi_q(p_n, F_{1,n})(y)$, where $F_{1,n}(y) = \frac{1}{n}\sum_{i=1}^n R_i I(Y_i \leqslant y)/p_n$ and $p_n = \frac{1}{n}\sum_{i=1}^n R_i$, Then, $\mu(F)$ is estimated by $\mu(F_n)$. This estimator is asymptotically equivalent to constructing an estimator of μ using the following α-specific procedure: (a) estimate f_1 by its empirical histogram, (b) plug this histogram into the right-hand side of (1) to compute an estimator of f_0 (see Figure 1), (c) compute an estimator of f by taking a average of the histograms in (a) and (b), weighted by the empirical proportion of completers, and (d) compute an estimator of μ by evaluating the mean associated with the histogram in (c).

It is straightforward to prove that $\sqrt{n}(\mu(F_n) - \mu(F))$ converges to a mean zero normal distribution with influence curve

$$IC_\mu(O; G_O) = \left\{RY - p\int_{-\infty}^{\infty} y\,\mathrm{d}F_1(y)\right\} - (R-p)\left\{\frac{\int_{-\infty}^{\infty} y\exp(q(y))\,\mathrm{d}F_1(y)}{\int_{-\infty}^{\infty}\exp(q(y))\,\mathrm{d}F_1(y)}\right\}$$
$$+ \frac{(1-p)}{p\int_{-\infty}^{\infty}\exp(q(y))\,\mathrm{d}F_1(y)}\left\{Y - \frac{\int_{-\infty}^{\infty} y\exp(q(y))\,\mathrm{d}F_1(y)}{\int_{-\infty}^{\infty}\exp(q(y))\,\mathrm{d}F_1(y)}\right\}\{R\exp(q(Y))\}$$

The asymptotic variance of $\mu(F_n)$ is then equal to $\sigma_\mu^2(G_O) = E[IC_\mu(O; G_O)^2]$ and it can be estimated by $\sigma_\mu^2(G_{O,n}) = E_n[IC_\mu(O; G_{O,n})^2]$, where $G_{O,n} = (p_n, F_{1,n})$ and $E_n[\cdot]$ is the empirical expectation operator.

5. Bayesian inference

To proceed with a Bayesian analysis, an expert must specify his/her prior beliefs about the model parameters. In light of the pattern-mixture model (1) and its selection model analog (3) where $q(Y) = \alpha\log(Y)$, we can parametrize the model by either (a) (α, p, F_1) or (b) (α, η, F). Our goal is to estimate the posterior distribution of $\mu = \mu(F)$. We denote this posterior distribution by $\pi(\mu|\boldsymbol{O})$. We know that the posterior for μ can be written as

$$\pi(\mu|\boldsymbol{O}) = \int \pi(\mu|\boldsymbol{O}, \alpha)\pi(\alpha|\boldsymbol{O})\,\mathrm{d}\alpha. \tag{1}$$

Furthermore, under parametrization (a), $\pi(\alpha|\boldsymbol{O})$ can be written as

$$\pi(\alpha|\boldsymbol{O}) = \int \pi(\alpha|p, F_1, \boldsymbol{O})\pi(p, F_1|\boldsymbol{O})\,\mathrm{d}\nu(p, F_1) \tag{2}$$

and under parametrization (b), $\pi(\alpha|\boldsymbol{O})$ can be written as

$$\pi(\alpha|\boldsymbol{O}) = \int \pi(\alpha|\eta, F, \boldsymbol{O})\pi(\eta, F|\boldsymbol{O})\,\mathrm{d}\nu(\eta, F).$$

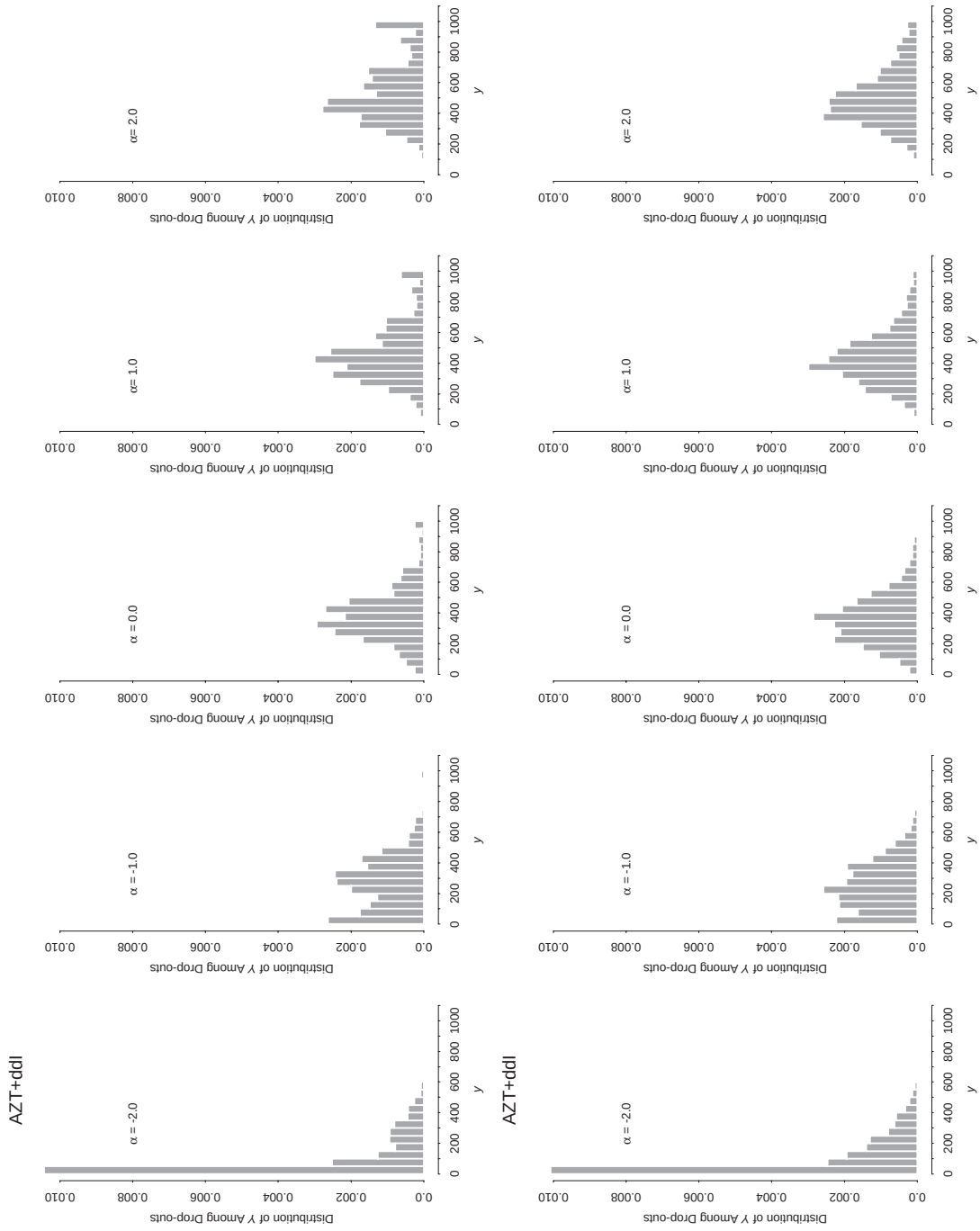

Fig. 1. Treatment-specific imputed distributions of CD4 count at week 56 (Y) for drop-outs as a function of the selection bias parameter α.

From (5), note that the posterior distribution for α will be exactly equal to the prior distribution on α (i.e. $\pi(\alpha|\boldsymbol{O}) = \pi(\alpha)$) if and only if the prior beliefs for α are independent of the prior beliefs about p and F_1 (i.e. $\pi(\alpha, p, F_1) = \pi(\alpha)\pi(p, F_1)$). We do not believe that such prior independence is substantively plausible based on our queries of three knowledgeable physicians. Specifically we asked these physicians whether seeing a higher than expected mean CD4 count in completers (a function of F_1) would influence their belief as to the magnitude of selection bias encoded in α. All agreed that a higher than expected mean observed CD4 count would indicate that there was probably a substantial degree of selection bias, with subjects with low CD4 counts being preferentially censored. Thus, to encode the beliefs of these experts, one would need to directly specify a dependent prior, which may be a difficult task. In contrast, the experts felt that α and (F, η) might be *a priori* independent. Thus, the second set of parameters, (α, η, F) will be the parameters on which we specify our priors. But, it is important to note that specifying such independence actually induces the desired dependence between the prior for α and (p, F_1). To see this, remember that in Section 4.1, we showed that for each α (i.e. $q(Y) = \alpha \log(Y)$), there is a one-to-one mapping between (p, F_1) and (η, F). That is, there exists a function $\Psi_\alpha(\cdot, \cdot)$, indexed by α, such that $(\eta, F) = \Psi_\alpha(p, F_1)$ and $(p, F_1) = \Psi_\alpha^{-1}(\eta, F)$. Thus, specifying independent priors for α and (η, F) induces a joint prior for $(\alpha, p, F_1) = (\alpha, \Psi_\alpha^{-1}(\eta, F))$ and α is not independent of $(p, F_1) = \Psi_\alpha^{-1}(p, F)$. This implies that the posterior for α will not be the same as the prior for α. In the next two sections, we discuss prior specification and a sampling algorithm to obtain the posterior for μ.

5.1 *Priors*

To complete the model, we need to specify prior distributions for α, η, and F. We specify independent priors for each:

$$\pi(\alpha, \eta, F) = \pi(\alpha)\pi(\eta)\pi(F).$$

In particular, we will assume an informative prior on α, a relatively non-informative prior on η, and a Dirichlet process prior on F, with known degrees of freedom, df (precision), and known base measure, $H(\boldsymbol{\theta})$, with unknown parameter vector, $\boldsymbol{\theta}$ (Antoniak, 1974). A hyper-prior on $\boldsymbol{\theta}$ must also specified. In our analysis of the ACTG 175 data, we took the base measure of the Dirichlet process to be log-normal with parameter vector $\boldsymbol{\theta} = (\gamma, \tau^2)$—the log-normal distribution is a common model for CD4 counts (see, for example, DeGruttola and Tu, 1994). By setting df small, we allow the prior dispersion around the base measure to be large. This is consistent with our view that it is rare to have strong prior beliefs about the distributional form of Y. Regardless of this view, however, our sampling scheme in the next section can be applied to situations where df is taken to be large and/or the prior on α is taken to be relatively non-informative.

To elicit a prior on α, one can ask an expert about the odds of drop-out for a proportional change in response. If the expert is comfortable with a parametric form for the prior, a most likely value might be specified along with several quantiles to uniquely determine the prior distribution. An alternative approach would be to use a non-parametric prior for which the expert would provide a best guess and then attach weights to intervals around that guess to form a histogram. The histogram might then be smoothed to facilitate inference. For a good discussion of elicitation of priors, see Chaloner (1996).

As a word of caution, we note that when informative priors on both α and F are chosen, we may observe situations where the priors, taken together, are not 'compatible' with the observed data. For example, suppose in our analysis of the ACTG 175 data, we assumed a left-skewed parametric family for F and assumed that α has most of its probability mass between -2.0 and -1.0. Given the observed data ($\alpha = 0$ in Figure 1), a sampling algorithm would like to impute large values of Y for the missing data based on the distributional form on F, while from the assumptions about α, the algorithm would like to impute

small values of Y for the missing data. This can result in bi-modal posterior distributions for $(\alpha, \eta, \mu(F))$. We believe such a scenario would be viewed as a problem and unsatisfactory for inferences. To avoid the bi-modality, we suggest that the experts re-evaluate their priors by reducing the 'informativeness' of one of them.

5.2 *Sampling from the posterior distribution*

To sample from the posterior distribution of $(\alpha, \eta, F, \boldsymbol{\theta})$, we will use a Gibbs sampling algorithm with Metropolis–Hastings steps (Smith and Roberts, 1993). We will also include a data augmentation step (Tanner and Wong, 1987) which will greatly simplify the algorithm by making it a complete data problem upon inclusion of the augmented data. Each of the K iterations of the algorithm will proceed as follows:

1. sample from $\pi(F, \boldsymbol{\theta}|\alpha, \eta, \boldsymbol{Y}_{miss}, \mathbf{O})$
2. sample from $\pi(\alpha, \eta|F, \boldsymbol{\theta}, \boldsymbol{Y}_{miss}, \mathbf{O})$
3. sample from $\pi(\boldsymbol{Y}_{miss}|\alpha, h, F, \boldsymbol{\theta}, \mathbf{O})$

where $\boldsymbol{Y}_{miss}$ is vector of outcomes for drop-outs. We now provide details on each step.

To sample from $\pi(F, \boldsymbol{\theta}|\alpha, \eta, \boldsymbol{Y}_{miss}, \mathbf{O})$ in Step 1, we will proceed in several substeps. We first note that if we place conjugate priors on $\boldsymbol{\theta}$, the full conditional distribution of each component will have known forms, after integrating out F (Doss, 1994). If we were to sample directly from $\pi(\boldsymbol{\theta}|F, \alpha, \eta, \boldsymbol{Y}_{miss}, \mathbf{O})$, the conditional distribution of $\boldsymbol{\theta}$ is not of closed form and we would need to use another Metropolis–Hastings algorithm and evaluate a discrete approximation to the Dirichlet process prior when computing the acceptance ratio. Since sampling from the conditional distribution of $\boldsymbol{\theta}$ with F integrated out is easier, we consider the following factorization:

$$\pi(F, \boldsymbol{\theta}|\alpha, \eta, \boldsymbol{Y}_{miss}, \mathbf{O}) = \pi(F, \boldsymbol{\theta}|\boldsymbol{Y}_{miss}, \mathbf{O}) = \pi(\boldsymbol{\theta}|\boldsymbol{Y}_{miss}, \mathbf{O})\pi(F|\boldsymbol{\theta}, \boldsymbol{Y}_{miss}, \mathbf{O}).$$

We propose to sample from the distributions of each component of $\boldsymbol{\theta}$ conditional on $(\boldsymbol{Y}_{miss}, \mathbf{O})$, which will have known forms. We do this 10–20 times to obtain a single sample point from the joint distribution of the $\boldsymbol{\theta}$. Then, conditional on $\boldsymbol{\theta}$, we sample from the full conditional distribution of F, with details below, to obtain a sample point from the joint full conditional of $(F, \boldsymbol{\theta})$. To sample from the full conditional of F, we need to do a discrete approximation to the infinite-dimensional distribution. The posterior of F will be a Dirichlet process prior with base measure a mixture of $H(\boldsymbol{\theta})$ and point masses at the observed and missing Y with weight $df/(df + n)$ associated with the former part of the mixture. The algorithm, adapted from Doss (1994), is as follows:

1a. Fix J to be a large integer
1b. Draw $B_1, \ldots, B_J$ i.i.d. from a $Beta(1, df + n)$
1c. Draw $V_1, \ldots, V_J$ i.i.d. using the following scheme. For each j, draw $U_j \sim$ Uniform(0, 1).
  1c.1 If $U_j < df/(df + n)$, draw V_j from $H(\boldsymbol{\theta})$.
  1c.2 Otherwise, draw V_j from the empirical cdf of Y (i.e. mass $1/n$ at each Y_i)
1d. Form $F_J = \sum_{j=1}^{J} P_j \delta_{V_j}$, where $P_j = B_j \prod_{r=1}^{j-1}(1 - B_r)$ and δ_a is the probability measure giving unit mass at the point a. This will be an approximate sample from the full conditional.

To sample from the full conditional distribution of (α, η) in Step 2, we will use a Metropolis–Hastings algorithm with candidate distribution a normal- or t-approximation to the full conditional distribution. With the augmented data, this is equivalent to sampling the coefficients in a Bayesian logistic regression model.

To sample from $\pi(\boldsymbol{Y}_{miss}|\alpha, \eta, F, \boldsymbol{\theta}, \mathbf{O})$ in Step 3, we can use the following approach for each missing value, $Y_{miss,}$:

3a. Draw an observation, Y_{cand}, from F, using an inverse cdf approach.
3b. Draw $U \sim \text{Uniform}(0, 1)$.
3c. If $U < \exp(\eta + \alpha \log(Y_{cand}))/(1 + \exp(\eta + \alpha \log(Y_{cand})))$, then set $Y_{miss} = Y_{cand}$. Otherwise, go to step 3a.

By 'imputing' the missing data in this way, we can work with the simpler complete data problem.

5.3 *Large sample approximation theory*

With $df << n$ and n large, a semiparametric version of the Bernstein–von Mises theorem (van der Vaart, 2000) suggests that the posterior of $\mu(F)$ given α will be well approximated by a normal distribution with mean $\mu(F_n; \alpha)$ and variance $\sigma^2_\mu(G_{O,n}; \alpha)/n$, where $\mu(F_n; \alpha)$ and $\sigma^2_\mu(G_{O,n}; \alpha)$ are the estimated mean and estimated asymptotic variance when $q(Y) = \alpha \log(Y)$. In addition, it can be shown that $\pi(\alpha|\boldsymbol{O}) = \pi(\alpha|G_O = G_{O,n}) + o_P(1)$. In light of (4), the posterior of $\mu(F)$ will then be well approximated by a mixture of a Normal$(\mu(F_n; \alpha), \sigma^2_\mu(G_{O,n}; \alpha)/n)$ over $\pi(\alpha|G_O = G_{O,n})$. Since $\pi(\alpha|G_O)$ is not closed form under our prior specifications, we need to use the sampling algorithm in the previous section.

When $df >> n$, n is large, and we specify relatively non-informative priors for α, η, $\boldsymbol{\theta}$, the model is approximately equivalent to a frequentist model in which (3) holds with $q(Y) = \alpha \log(Y)$, α unknown, $Y \sim H(\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ unknown. Assuming $H(\boldsymbol{\theta})$ induces enough restrictions, all of the parameters in the frequentist model will be well identified. Thus, a parametric version of the Bernstein–von Mises theorem tells us that the posterior of $\mu(F)$ is well approximated by the distribution of the maximum likelihood estimator for the mean of Y in the frequentist model.

5.4 *Checking convergence*

When df is small and the sample size n is large, we saw in the previous section that given α and the data $\boldsymbol{O}$, the distribution of $\mu(F)$ should be approximately normal with mean $\mu(F_n; \alpha)$ and variance $\sigma^2_\mu(G_{O,n}; \alpha)/n$. Now, our sampling scheme produces K draws from the joint posterior of α and F, $(\alpha^{(1)}, F^{(1)}), \ldots, (\alpha^{(K)}, F^{(K)})$. Thus, given α, we can check the convergence of our posterior sampling scheme by plotting a smoothed histogram of the $\mu(F^{(k)})$, for which the associated $\alpha^{(k)}$ fall within a small interval around α. If convergence is reached, then this histogram should be approximately normal with mean $\mu(F_n; \alpha)$ and variance $\sigma^2_\mu(G_{O,n}; \alpha)/n$. This can be repeated for various α.

As a more generic check, we suggest the multiple chains approach of Gelman and Rubin (1992). This approach will also help diagnose bi-modality of the posterior distribution.

5.5 *Model checking*

The fit of the model to the observed data is the only way to assess the adequacy of the model. This is especially important when df is taken to be large. A simple way to check the model here is to see how well the posterior distribution of F_1 matches up with the empirical cdf of the observed data. We can use the posterior predictive checking approach of Gelman, Meng, and Stern (1996). That is, we sample a new set of observed Y from their posterior predictive distribution, using the same approach that we used to sample $\boldsymbol{Y}_{miss}$. For each sample, we compute an empirical cdf based on the new set of observed Y and plot. By overlaying the empirical cdf of the actual observed data, we can see whether it is consistent with what the model is predicting.

6. ANALYSIS OF ACTG 175

6.1 *Parametric selection model analysis*

One approach that has been proposed in the statistical literature is to specify parametric models for both the conditional distributional of R given Y and the marginal distribution of Y (Diggle and Kenward, 1994). For ACTG 175, it might be assumed that Y is log-normal with parameter vector $\boldsymbol{\theta} = (\gamma, \tau^2)$ and the conditional distribution of R given Y follows (3) with $q(Y) = \alpha \log(Y)$. Due to the log-normal assumption, α is identified. Inference could then proceed by maximum likelihood. By letting df be very large and assuming relatively non-informative priors on α, η, and $\boldsymbol{\theta}$, we know that the posterior distributions of α and $\mu = \exp(\gamma + \tau^2/2)$ will approximate the marginal distributions of their maximum likelihood estimators. We performed such an analysis by letting $df = 10\,000$, $\pi(\alpha) \sim N(0, 10.0)$, $\pi(\eta) \sim N(0, 100)$, $\pi(\gamma) \sim N(5.5, 1.0)$ and $\pi(1/\tau^2) \sim G(1, 10)$, where $N(a, b)$ denotes a normal distribution with mean a and variance b and $G(c, d)$ denotes a gamma distribution with scale c, shape d, mean cd and variance cd^2. The prior mean for γ was chosen based on the observed overall mean of log CD4 cells at baseline and the fact that the overall health of the cohort was expected to decline over the study period; the prior mean of τ^2 was set equal to the inverse of the overall variance of log CD4 cells at baseline; the prior variances of γ and τ^2 were chosen large enough so that the data and modeling assumptions would ultimately determine the posterior distributions of γ and τ^2.

In Figure 2, we present the treatment-specific posterior distributions of α and μ (dashed lines), based on $K = 10\,000$ iterations (first 1000 iterations discarded; overall acceptance rate in Step 2 of the algorithm was 72.6% and 74.8% in the AZT+ddI and ddI arms, respectively). The approximate maximum likelihood estimates (standard errors; 95% credible intervals) for α are -2.58 (0.24; $[-3.00, -2.09]$) and -2.76 (0.26; $[-3.25, -2.22]$) for the AZT+ddI and ddI arms, respectively. Under the log-normality assumption, we would reject the treatment-specific null hypothesis of missing at random. The corresponding estimates for μ are 303.42 (13.63;[277.67,331.20]) and 296.68 (13.51;[271.49,324.17]). The posterior distribution of the difference between the means in the AZT+ddI and ddI arms is displayed in Figure 3. The approximate maximum likelihood estimate of the difference is 6.74 (19.27; $[-30.89, 43.92]$), suggesting no significant treatment effect. To check the model fit, we use the posterior predictive checking approach described in Section 5.5. The first row of Figure 4 displays the empirical distribution of the observed CD4 counts at week 56 (solid line) along with 100 empirical distribution functions of observed outcomes drawn from its posterior predictive distribution. As the the figure illustrates, the model does not fit the observed data well. To achieve a better fit, we tried two alternative models for the marginal distribution of Y. In particular, we let the cube root and square root of Y be normally distributed. These models fit slightly better than the log-normal, but still did not fit the observed data well. Under these alternative models, the approximate maximum likelihood estimates and confidence intervals for α were of similar order of magnitude as the log-normal model.

6.2 *Frequentist non-parametric sensitivity analysis*

When the distribution of the CD4 count at week 56 is left unspecified, we saw in Section 4 that α is not identified. In the absence of such distributional information, the best that can be achieved from a frequentist perspective is to perform a sensitivity analysis. In Figure 5, we present the treatment-specific estimated overall mean CD4 count at week 56 (with 95% confidence intervals) as a smooth function of the selection bias parameter α (α ranges from -2.0 to 2.0). Note how the sampling variability (as indicated by the width of the confidence intervals) is dominated by the uncertainty in α. That is, the width of any given confidence interval is approximately 40–50 CD4 cell counts, while the range of estimated means across α is approximately 200–300 CD4 cell counts. Common statistical practice is to simply report one of these confidence intervals which can grossly misrepresent treatment efficacy.

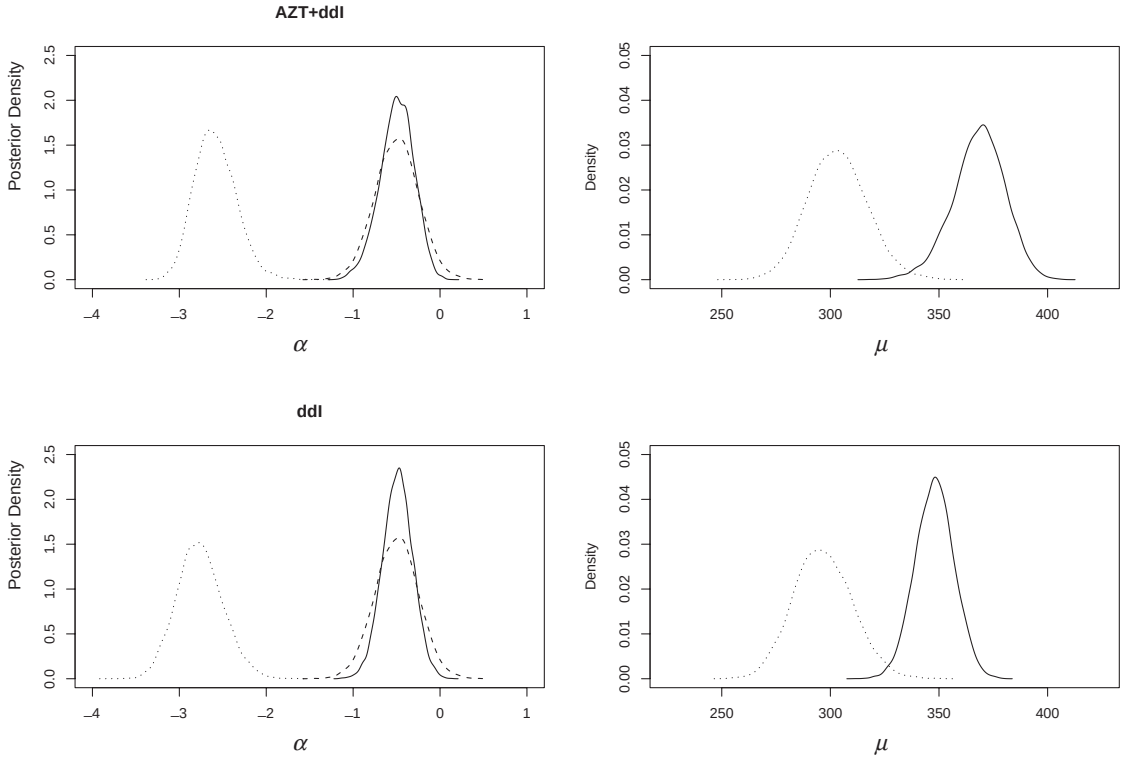

Fig. 2. Treatment-specific posterior distributions for α and μ. Dashed lines (Model with $df = 10,000$, $\pi(\alpha) \sim N(0, 10)$, $\pi(\eta) \sim N(0, 100)$, $\pi(\gamma) \sim N(5.5, 1.0)$ and $\pi(1/\tau^2) \sim G(1, 10)$); Solid lines (Model with $df = 1$, $\pi(\alpha) \sim N(-0.5, 0.25^2)$, $\pi(\eta) \sim N(0, 100)$, $\pi(\gamma) \sim N(5.5, 1.0)$ and $\pi(1/\tau^2) \sim G(1, 10)$). Dotted line (Prior $\pi(\alpha) \sim N(-0.5, 0.25^2)$).

In Figure 6, we present a contour plot of the Z-statistic (estimated difference in means divided by standard error of the difference) associated with the test of the null hypothesis of no treatment difference as a function of treatment-specific selection bias parameters. On the horizontal (vertical) axis, we vary the selection bias parameter for the AZT+ddI (ddI) arm. Regions marked with a treatment label indicate that, for selection bias parameter combinations in the region, a 0.05 level test of the null hypothesis would be rejected in favor of that treatment. The solid point in the contour plot indicates the result from the missing at random assumption in both treatment groups. The conclusion from this contour plot is that with mild levels of differential selection bias in the treatment arms (e.g. $\alpha = 0$ in the ddI arm and $\alpha = -0.025$ in the AZT+ddI arm), we would change the conclusion based on the default analysis. As a result, the evidence in favor of AZT+ddI appears to be 'weaker' than that based on missing at random. As we see, this sensitivity analysis may be viewed as limited in its use since it does not provide an overall quantification of the strength of evidence, accounting for uncertainty in beliefs about selection bias.

6.3 *Flexible Bayesian analysis*

When a decision is required, the flexible Bayesian methodology described in Section 5 can be used to summarize the treatment efficacy in the presence of the uncertainly regarding the distribution of the outcome and the level of selection bias.

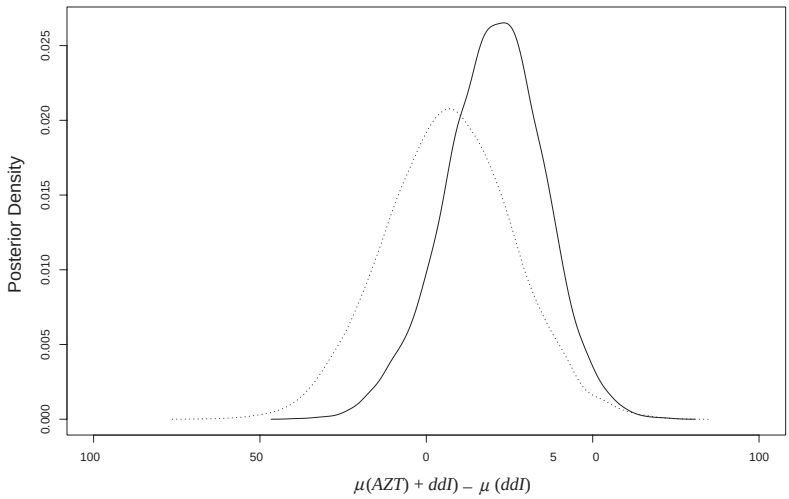

Fig. 3. Posterior distribution for $\mu(AZT + ddI) - \mu(ddI)$. Dashed lines (Model with $df = 10,000$, $\pi(\alpha) \sim N(0, 10)$, $\pi(\eta) \sim N(0, 100)$, $\pi(\gamma) \sim N(5.5, 1.0)$ and $\pi(1/\tau^2) \sim G(1, 10)$); Solid lines (Model with $df = 1$, $\pi(\alpha) \sim N(-0.5, 0.25^2)$, $\pi(\eta) \sim N(0, 100)$, $\pi(\gamma) \sim N(5.5, 1.0)$ and $\pi(1/\tau^2) \sim G(1, 10)$)

For each treatment group, we assumed that the distribution of Y, F, followed a Dirichlet process mixture prior with precision $df = 1$, and log-normal base measure with parameter vector $\theta = (\gamma, \tau^2)$. We assumed that the hyper-priors for γ and τ^2 were independent. In particular, we let $\pi(\gamma) \sim N(5.5, 1)$, $\pi(1/\tau^2) \sim \Gamma(1, 10)$. In addition, we assumed an independent, non-informative, normal prior on η, i.e. $\pi(\eta) \sim N(0, 100)$. For α, we use $\pi(\alpha) \sim N(-0.5, 0.25^2)$. That is, the prior belief is that subjects with lower CD4 counts (under full compliance) at week 56 are more likely to drop out. Specifically, the prior states that there is a 95% chance that the odds ratio of drop-out for subjects with a two-fold change in CD4 cells at week 56 is between 1 and 2, with a most probable value around 1.4.

Figure 2 displays the treatment-specific posterior distributions of α and μ (solid lines), based on $K = 10\,000$ iterations (first 1000 iterations discarded; acceptance rate in Step 2 of the algorithm was 80.0% and 80.3% in the AZT+ddI and ddI arms, respectively). The prior distribution of α is the dotted line and demonstrates the difference between the prior and posteriors. The posterior means (95% credible intervals) for α are -0.50 ($[-0.90, -0.14]$) and -0.49 ($[-0.83, -0.15]$) for the AZT+ddI and ddI arms, respectively. For comparison, the prior mean (95% credible interval) for α was -0.50 ($[-1.0, 0.0]$) for both treatment arms. The treatment-specific posterior distributions of α are tighter than the prior distributions, indicating a weak level of induced *a priori* dependence between α and (p, F_1). The posterior means (95% credible intervals) for μ are 368.24 ([342.43, 390.57]) and 348.00 ([330.40, 365.40]) for the AZT+ddI and ddI arms, respectively. Figure 3 displays the posterior distribution of the difference between $\mu(AZT + ddI)$ and $\mu(DDI)$ (solid line). The posterior mean (95% credible interval) is 20.24 ($[-11.12, 48.63]$). The posterior probability that $\mu(AZT + ddI)$ is greater than $\mu(DDI)$ is 90.76%. After accounting for prior beliefs regarding selection bias, there appears to be relatively strong evidence in favor of combination therapy.

In Figure 7, we display the treatment-specific convergence diagnostic described in Section 5.4 for $\alpha = -0.8, -0.65, -0.5, -0.35, -0.2$. The solid line is the estimated distribution of $\pi(\mu|\boldsymbol{O}, \alpha)$ based on our non-parametric maximum likelihood results of Section 4 and the dotted line is the estimated distribution based on the Gibbs sampling scheme. The two estimates are quite close, providing support for convergence. The second row of Figure 4 displays the empirical distribution of the observed CD4 counts

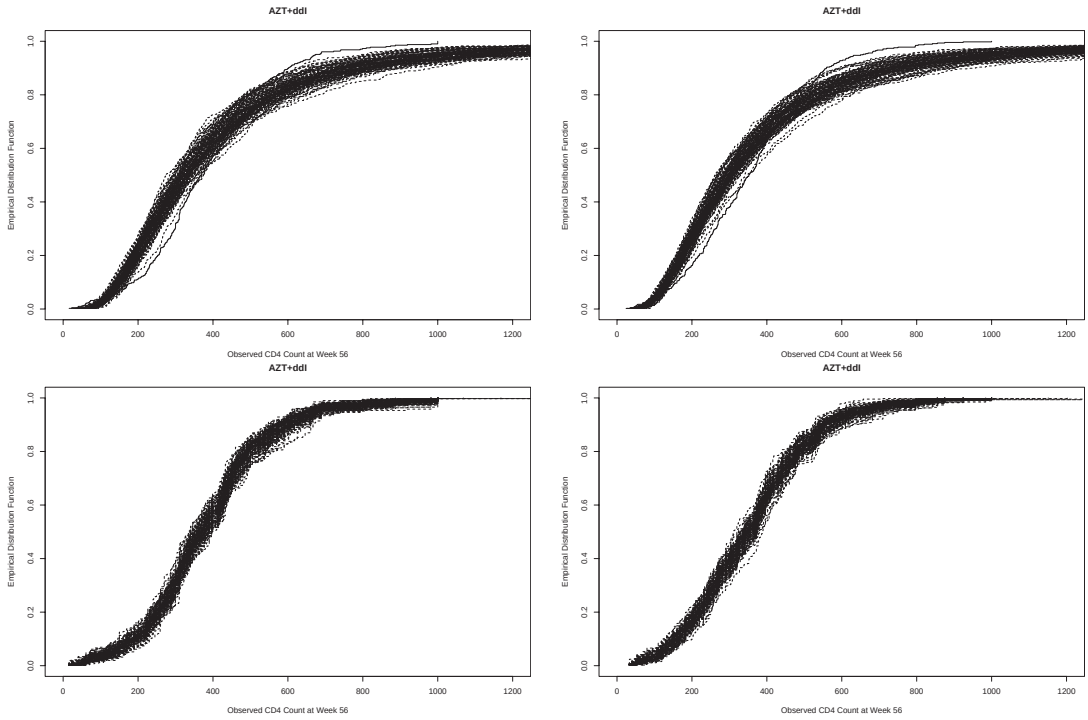

Fig. 4. Model checking: empirical distribution function of observed CD4 counts at week 56 (solid line) versus 100 draws from the posterior predictive distribution of the empirical cumulative distribution function of observed CD4 counts at week 56 (dashed lines). The first row corresponds to the fully parametric selection model and the second row corresponds to the flexible semiparametric selection model.

at week 56 (solid line) along with 100 empirical distribution functions of observed outcomes drawn from its posterior predictive distribution. As the the figure illustrates, the model, as expected, fits the observed data well.

6.4 *Summary and comparison of approaches*

In evaluating treatment effects in the presence of missing data, the analyst usually starts with the default missing at random analysis. Under missing at random, the estimated means are 385 and 360 in the AZT+ddI and ddI arms, respectively. The difference in means is 25 CD4 cells and the null hypothesis of no treatment difference is rejected.

Recognizing that missing at random is likely to fail, the analyst might consider a parametric selection model analysis, similar to the one conducted in Section 6.1. There are two important lessons to be learned from this analysis. The first lesson is that model selection is difficult and model checking is critical. In our analysis, we found that some of the models typically used to fit CD4 count data did not fit the observed data well. To achieve better fits, one needs to either use a more flexible model for the distribution of Y or fit a more flexible form for the selection bias function $q(Y)$. The second lesson is that the distributional form of Y can determine the magnitude of selection bias and one must make sure that the level of selection bias is substantively plausible. Using the log-normal assumption, the estimated value of α is -2.58 and -2.76 in the AZT+ddI and ddI arms, respectively. Missing at random is rejected in both treatment arms. These levels of selection bias are enormous; the imputed distribution of Y among drop-outs is more extreme

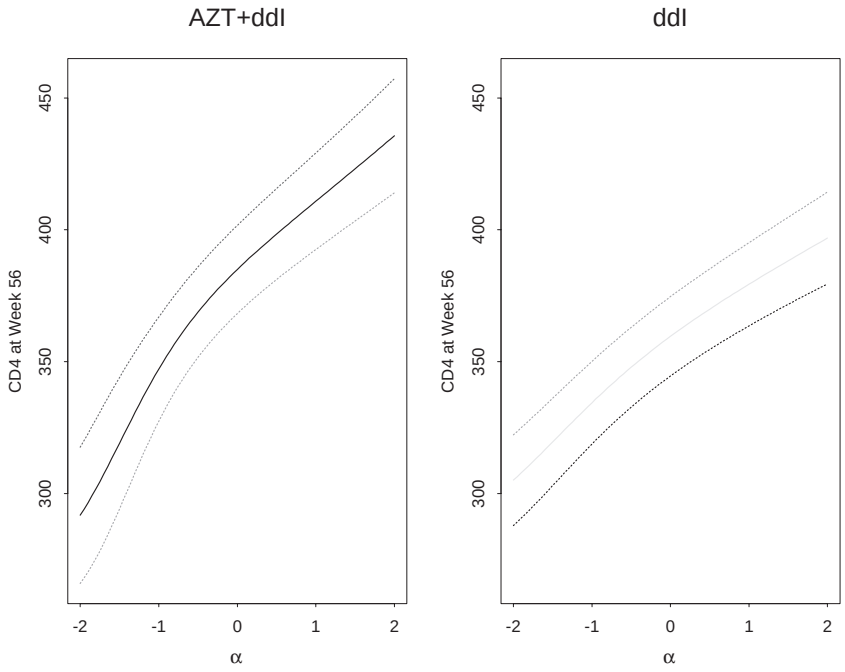

Fig. 5. Treatment-specific estimated mean CD4 count at week 56 (with 95% confidence intervals) as a function of the selection bias parameter α.

than the left-most histograms in Figure 1. They correspond to the belief that for two subjects who have a two-fold difference in CD4 cells at week 56, the sicker subject is 6.0 and 6.7 times as likely to drop out in the AZT+ddI and ddI arms, respectively. These levels of selection bias are highly implausible. With these levels, the estimated means are 303 and 297 for the AZT+ddI and ddI arms. This is a huge reduction from the estimated means under missing at random. The difference in means is 6.7 CD4 cells and is not statistically significant. Similar results were observed in the alternative models we considered. While one can argue that these levels of selection bias are due to ill-fitting models, we conjecture that one can posit a more flexible model for the marginal distribution of Y which fits the data well but nevertheless still identifies a highly implausible level of selection bias.

The frequentist non-parametric analysis in Section 6.2 suggested that evidence in favor of AZT+ddI is weaker than that provided by the missing at random analysis. However, the drawback of the sensitivity analysis is that a single answer is not provided and the level of uncertainty is not quantified. Using treatment-specific informative priors on α, the flexible Bayesian analysis of Section 6.3 provides a quantification of uncertainty through posterior distributions. In this analysis, the treatment-specific distributions for α are slightly narrower than the prior specifications, indicating that, within the context of our fully Bayesian model, the data provide relatively little information about α. The estimated means are 368 and 348 in the AZT+ddI and ddI arms, which are, as expected, lower than the missing at random means, and more plausible than those from the fully parametric analysis. The posterior distributions (solid lines in Figure 3) indicate the degree of uncertainty regarding the treatment-specific means. The degree of uncertainty is comparable to that provided by parametric analysis and much larger than that of the missing at random analysis. The estimated mean difference is 20 CD4 cells and the posterior distribution for the difference (solid line in Figure 4) indicates the level of uncertainty as reflected by the span of the 95%

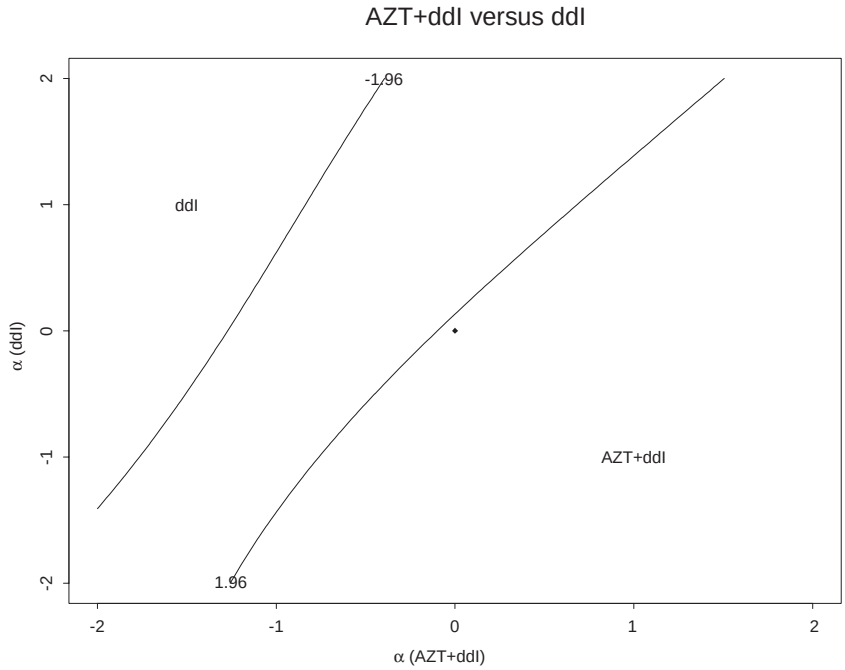

Fig. 6. Contour plot of the Z-statistic (estimated difference in means divided by standard error of the difference) associated with the test of the null hypothesis of no treatment difference as a function of treatment-specific selection bias parameters.

credible interval and the 91% chance that AZT+ddI is superior to ddI. This is a concise representation of the strength of evidence regarding the treatment effect. As this analysis incorporates prior beliefs about selection bias and fits the observed data well, it may be more plausible than the missing at random analysis.

7. Discussion

In our view, the fully parametric approach above should only be used when there is strong scientific evidence to support the use of particular parametric models for the distribution of the outcome *and* the selection mechanism. When there are only strong prior beliefs about the distribution of the outcome, the approach of Lee and Berger (2001) or our approach with large weight given to the base measure of the prior distribution for the outcome and a high-dimensional parametrization of the selection bias function with vague priors can be used. In situations where informative prior beliefs about the selection bias can be quantified, our flexible Bayesian approach is an attractive way of summarizing the treatment effect and its associated uncertainty. If a flexible Bayesian analysis is implemented, we feel that it should be conducted in conjunction with the sensitivity analysis, as this latter analysis provides informal and formal checks for the former. Finally, if experts cannot agree on the distributional form or on the nature of selection bias, then the only objective analysis is to present worst-case bounds.

In the ACTG 175 data, missingness of the outcome occurred for multiple reasons (i.e. non-compliance, skipped clinic visits, and loss to follow-up). In our selection model, we did not distinguish between these various causes of missingness, which could very well have different relationships with the outcome of interest. It is not difficult to extend our approach to this setting. Specifically, one could re-define R to have

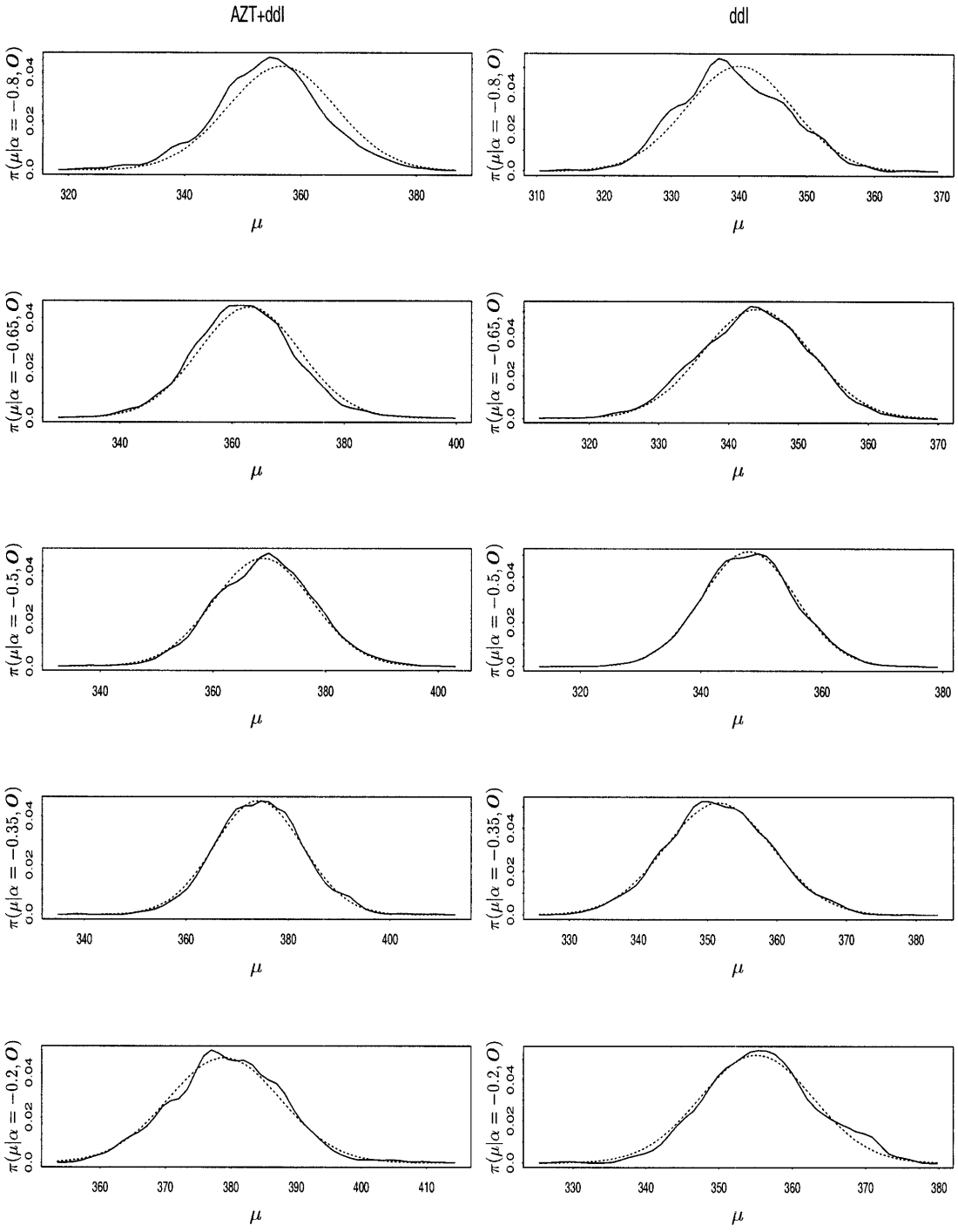

Fig. 7. Treatment-specific convergence diagnostic.

unique values corresponding to each major type of missingness and a unique value for completion. Then, one would fit a polytomous logistic regression with type-specific intercepts and selection bias functions. To allow for differential relationships, these functions would have to be parameterized separately in order to elicit priors.

In most randomized trials, the data structure is much more complicated than the one handled in this paper. Specifically, baseline information is usually collected and the primary outcomes may be failure times or repeated measures. The sensitivity analysis ideas have been extended to deal with these more realistic settings. In future work, we will focus on the extensions of our flexible Bayesian approach. The task will be considerably more difficult, as more information implies more modeling and greater prior specifications. For example, if we were to consider high-dimensional baseline prognostic factors X as part of our observed data, then the analog of our model (3) would be, say,

$$\text{logit } P[R = 0|X, Y] = h(X) + q(X, Y)$$

where $h(X)$ is some unknown function of X and $q(x, y)$ is a specified function of x and y. We would then need to specify an informative prior for q, and flexible priors for $h(X)$ and the conditional distribution of Y given X. Flexible priors are important here to prevent identification of q in ways that are unintended. In addressing even this simple extension, substantial dimension reduction will be required and the computational complexity of the sampling algorithm will increase substantially.

Future work will also focus on developing (1) techniques for elicitation of prior beliefs for the selection bias parameter, (2) faster sampling algorithms, and (3) formal proofs of the large sample properties of our flexible Bayesian procedure.

ACKNOWLEDGEMENTS

This research was partially supported by National Institute of Health grants CA85295, MH56639, HD38209, A132475. The authors would like to thank the associate editor and two reviewers for their helpful comments, which have greatly improved the quality of the manuscript.

REFERENCES

ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* **2**, 1152–1174.

BAYARRI, M. AND BERGER, J. (1998). Robust Bayesian analysis of selection models. *The Annals of Statistics* **26**, 645–659.

BAYARRI, M. AND DEGROOT, M. (1987). Bayesian analysis of selection models. *The Statistician* **36**, 137–146.

CHALONER, K. (1996). Elicitation of prior distributions. Berry, D. and Stangl, D. (eds), *Bayesian Biostatistics*. New York: Dekker, pp. 141–156.

DANIELS, M. AND HOGAN, J. (2000). Reparameterizing the pattern-mixture model for sensitivity analyses under informative drop-out. *Biometrics* **56**, 1241–1248.

DEGRUTOLLA, V. AND TU, X. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival. *Biometrics* **50**, 1003–1014.

DIGGLE, P. AND KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis (disc: P73-93). *Applied Statistics* **43**, 49–73.

DOSS, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *The Annals of Statistics* **22**, 1763–1786.

FORSTER, J. J. AND SMITH, P. W. F. (1998). Model-based inference for categorical survey data subject to non-ignorable non-response. *Journal of the Royal Statistical Society, Series B* **60**, 57–70.

GELMAN, A., MENG, X.-L. AND STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (disc: P760-807). *Statistica Sinica* **6**, 733–760.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (disc: P483-501, 503-511). *Statistical Science* **7**, 457–472.

HAMMER, S. M., KATZENSTEIN, D. A., HUGHES, M. D., GUN DACKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M., *et al.* (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine* **335**, 1081–1090.

LEE, J. AND BERGER, J. (2001). Semiparametric Bayesain analysis of selection models. *Journal of the American Statistical Association* **96**, 1397–1409.

LITTLE, R. (1994). A class of pattern-mixture models for normal missing data. *Biometrika* **81**, 471–483.

NEAL, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **6**, 353–366.

RAAB, G. M. AND DONNELLY, C. A. (1999). Information on sexual behavior when some data are missing. *Applied Statistics* **48**, 117–133.

ROBINS, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. Sechrest, L., Freeman, H. and Mulley, A. (eds), *Health Service Research Methodology: A Focus on AIDS*. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research, pp. 113–159.

ROBINS, J. M., ROTNITZKY, A. AND SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. Halloran, E. M. and Berry, D. (eds), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer, pp. 1–94.

ROTNITZKY, A., ROBINS, J. M. AND SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

ROTNITZKY, A., SCHARFSTEIN, D., SU, T. AND ROBINS, J. (2001). Methods for conducting sensitivity analysis of trials with potentially non-ignorable competing causes of censoring. *Biometrics* **57**, 103–113.

RUBIN, D. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **72**, 538–543.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **72**, 581–590.

SCHARFSTEIN, D., ROBINS, J., EDDINGS, W. AND ROTNITZKY, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event outcomes. *Biometrics* **57**, 404–413.

SCHARFSTEIN, D. O., ROTNITZKY, A. AND ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94**, 1096–1146.

SMITH, A. F. M. AND ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (disc: P53-102). *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 3–23.

TANNER, M. A. AND WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–540.

VAN DER VAART, A. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[*Received October 3, 2001; first revision July 16, 2002; second revision November 7, 2002; accepted for publication December 12, 2002*]