

## Abstract

Data with missing responses generated by a non-ignorable missingness mechanism can be analysed by jointly modelling the response and a binary variable indicating whether the response is observed or missing. Using a selection model factorisation, the resulting joint model consists of a model of interest and a model of missingness. In the case of non-ignorable missingness, model choice is difficult because the assumptions about the missingness model are never verifiable from the data at hand. For complete data, the Deviance Information Criterion (DIC) is routinely used for Bayesian model comparison. However, when an analysis includes missing data, DIC can be constructed in different ways and its use and interpretation are not straightforward. In this paper, we present a strategy for comparing selection models by combining information from two measures taken from different constructions of the DIC. A DIC based on the observed data likelihood is used to compare joint models with different models of interest but the same model of missingness, and a comparison of models with the same model of interest but different models of missingness is carried out using the model of missingness part of a conditional DIC. Our strategy is illustrated by an example with simulated missingness and an application which compares three treatments for depression using data from a clinical trial. We also examine issues relating to the calculation of the DIC based on the observed data likelihood.

## 1 Introduction

Missing data is pervasive in many areas of scientific research, and can lead to biased or inefficient inference if ignored or handled inappropriately. A variety of approaches have been proposed for analysing such data, and their appropriateness depends on the type of missing data and the mechanism that led to the missing values. Here, we are concerned with analysing data with missing responses thought to be generated by a non-ignorable missingness mechanism. In these circumstances, a recommended approach is to jointly model the response and a binary variable indicating whether the response is observed or missing. Several factorisations of the joint model are available, including the selection model factorisation and the pattern-mixture factorisation, and their pros and cons have been widely discussed (Kenward and Molenberghs, 1999; Michiels *et al.*, 2002; Fitzmaurice, 2003). In this paper, attention is restricted to selection models with a Bayesian formulation.

Spiegelhalter *et al.* (2002) proposed a *Deviance Information Criterion*, DIC, as a Bayesian measure of model fit that is penalised for complexity. This can be used to compare models in a similar way to the Akaike Information Criterion (for non-hierarchical models with vague priors on all parameters,  $DIC \approx AIC$ ), with the model taking the smallest value of DIC being preferred. However, for complex models, the likelihood, which underpins DIC, is not uniquely defined, but depends on what is considered as forming the likelihood and what as forming the prior. With missing data, there is also the question of what is to be included in the likelihood term, just the observed data or the missing data as well. For models allowing non-ignorable missing data, we must take account of the missing data mechanism in addition to dealing with the complication of not observing the full data.

Celeux *et al.* (2006) assess different DIC constructions for missing data models, in the context of mixtures of distributions and random effects models. Daniels and Hogan (2008), Chapter 8, discuss two different constructions for selection models, one based on the observed data likelihood,  $DIC_O$ , and the other based on the full data likelihood,  $DIC_F$ . However,  $DIC_F$  has proved difficult to implement in practice. The purpose of this paper is to first examine issues of implementation and usability of

$DIC_O$  and to clarify possible misuse. We then build on this to show how insights from  $DIC_O$  can be complemented by information from part of an alternative, ‘conditional’, DIC construction, thus providing the key elements of a strategy for comparing selection models.

In Section 2, we introduce selection models and review the general definition of DIC, before discussing how  $DIC_O$  and a DIC based on a likelihood that is conditional on the missing data,  $DIC_C$ , can provide complementary information about the comparative fit of a set of models. Issues concerning the calculation of  $DIC_O$  are discussed in Section 3, including choice of algorithm, plug-ins and sample size. In Sections 4 and 5 we describe the use of a combination of  $DIC_O$  and  $DIC_C$  to compare models for simulated and real data missingness respectively, before concluding with a discussion in Section 6.

## 2 DIC for selection models

We start this section by introducing the selection model factorisation, then discuss the general formula for DIC, and finally look at different constructions of DIC for selection models.

### 2.1 Introduction to selection models

Suppose our data consists of a univariate response with missing values,  $\mathbf{y} = (y_i)$ , and a vector of fully observed covariates,  $\mathbf{x} = (x_{1i}, \dots, x_{pi})$ , for  $i = 1, \dots, n$  individuals, and let  $\boldsymbol{\lambda}$  denote the unknown parameters of our model of interest.  $\mathbf{y}$  can be partitioned into observed,  $\mathbf{y}_{obs}$ , and missing,  $\mathbf{y}_{mis}$ , values, i.e.  $\mathbf{y} = (\mathbf{y}_{obs}, \mathbf{y}_{mis})$ . Now define  $\mathbf{m} = (m_i)$  to be a binary indicator variable such that

$$m_i = \begin{cases} 1: & y_i \text{ observed} \\ 0: & y_i \text{ missing} \end{cases}$$

and let  $\boldsymbol{\theta}$  denote the unknown parameters of the missingness function. The joint distribution of the full data,  $(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}|\boldsymbol{\lambda}, \boldsymbol{\theta})$ , can be factorised as

$$f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m}|\boldsymbol{\lambda}, \boldsymbol{\theta}) = f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\lambda}) \quad (1)$$

suppressing the dependence on the covariates, and assuming that  $\mathbf{m}|\mathbf{y}, \boldsymbol{\theta}$  is conditionally independent of  $\boldsymbol{\lambda}$ , and  $\mathbf{y}|\boldsymbol{\lambda}$  is conditionally independent of  $\boldsymbol{\theta}$ , which is usually reasonable in practice. This factorisation of the joint distribution is known as a selection model (Schafer and Graham, 2002). Both parts of the model involve  $\mathbf{y}_{mis}$ , so they must be fitted jointly. Consequently assumptions concerning the model of interest will influence the model of missingness parameters through  $\mathbf{y}_{mis}$ , and vice versa.

### 2.2 Introduction to DIC

*Deviance* is a measure of overall fit of a model, defined as -2 times the log-likelihood,  $D(\boldsymbol{\phi}) = -2\log L(\boldsymbol{\phi}|\mathbf{y})$ , with larger values indicating poorer fit. In Bayesian statistics deviance can be summarised in different ways, with the posterior mean of the deviance,  $\overline{D(\boldsymbol{\phi})} = E\{D(\boldsymbol{\phi})|\mathbf{y}\}$ , suggested as a sensible Bayesian measure of fit (Dempster, 1973) (reprinted as Dempster (1997)), though this is not penalised for model complexity. Alternatively, the deviance can be calculated using a point estimate

such as the posterior means for  $\phi$ ,  $D(\bar{\phi}) = D\{E(\phi|\mathbf{y})\}$ . In general we use the notation  $E(h(\phi)|\mathbf{y})$  to denote the expectation of  $h(\phi)$  with respect to the posterior distribution of  $\phi|\mathbf{y}$ . However, in more complex formula, we will occasionally use the alternative notation,  $E_{\phi|\mathbf{y}}(h(\phi))$ .

Spiegelhalter *et al.* (2002) proposed that the difference between these two measures,  $p_D = \overline{D(\phi)} - D(\bar{\phi})$ , is an estimate of the ‘effective number of parameters’ in the model. The DIC proposed by Spiegelhalter *et al.* (2002) adds  $p_D$  to the posterior mean deviance, giving a measure of fit that is penalised for complexity,

$$\text{DIC} = \overline{D(\phi)} + p_D. \quad (2)$$

DIC can also be written as a function of the log likelihood, i.e.

$$\text{DIC} = 2\log L\{E(\phi|\mathbf{y})|\mathbf{y}\} - 4E_{\phi|\mathbf{y}}\{\log L(\phi|\mathbf{y})\}. \quad (3)$$

More generally, if  $\bar{D}$  denotes the posterior mean of the deviance and  $\hat{D}$  denotes the deviance calculated using some point estimate, then  $\text{DIC} = 2\bar{D} - \hat{D}$ . We will refer to  $\hat{D}$  as a *plug-in deviance*, and the point estimates of the parameters used in its estimation as *plug-ins*. The value of DIC is dependent on the choice of plug-in estimator. The posterior mean, which is a common choice, leads to a lack of invariance to transformations of the parameters (Spiegelhalter *et al.*, 2002), and the reasonableness of the choice of the posterior mean depends on the approximate normality of the parameter’s posterior distribution. Alternatives to the posterior mean include the posterior median, which was investigated at some length by Spiegelhalter *et al.* (2002), and the posterior mode, which was considered as an alternative by Celeux *et al.* (2006).

Further, in hierarchical models we can define the prior and likelihood in different ways depending on the quantities of interest, which will affect the calculation of both  $\bar{D}$  and  $\hat{D}$  and hence DIC. The chosen separation of the joint density into prior and likelihood determines what Spiegelhalter *et al.* (2002) refer to as the *focus* of the model.

For complete data, DIC is routinely used by Bayesian statisticians to compare models, a practice facilitated by its automatic calculation by the WinBUGS software, which allows Bayesian analysis of complex statistical models using Markov chain Monte Carlo (MCMC) techniques (Spiegelhalter *et al.*, 2003). WinBUGS calculates DIC, taking  $\overline{D(\phi)}$  to be the posterior mean of  $-2\log L(\phi|\mathbf{y})$ , and evaluating  $D(\bar{\phi})$  as -2 times the log-likelihood at the posterior mean of the stochastic nodes. However, other values of DIC can be obtained by using different plug-ins.

When data include missing values, the possible variations in defining DIC are further increased. Different treatments of the missing data lead to different specifications, and there is also the question of what is to be included in the likelihood, just the observed data or the missing data as well.

### 2.3 Conditional DIC

One option is a conditional DIC, which treats the missing data as additional parameters (Celeux *et al.*, 2006). This can be written as:

$$\text{DIC}_C = 2\log L\{E(\lambda, \theta, \mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{m})|\mathbf{y}_{obs}, \mathbf{m}\} - 4E_{\lambda, \theta, \mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{m}}\{\log L(\lambda, \theta, \mathbf{y}_{mis}|\mathbf{y}_{obs}, \mathbf{m})\}.$$

For selection models the likelihood on which this is based is

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}) &\propto f(\mathbf{y}_{obs}, \mathbf{m} | \boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis}) \\ &= f(\mathbf{y}_{obs} | \boldsymbol{\lambda}) f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}). \end{aligned} \quad (4)$$

So the part of the likelihood calculated for the model of interest,  $f(\mathbf{y}_{obs} | \boldsymbol{\lambda})$ , is based on  $\mathbf{y}_{obs}$  only, and the part for the model of missingness,  $f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$ , is conditional on  $\mathbf{y}_{mis}$ . The plug-ins for  $DIC_C$  include the missing data,  $\mathbf{y}_{mis}$ , and are evaluated as  $E(\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})$ . However, conditional DICs have asymptotic and coherency difficulties (Celeux *et al.*, 2006; Little and Rubin, 1983).

The DIC automatically generated by WinBUGS in the presence of missing data is a conditional DIC, and WinBUGS produces DIC values for the model of interest and model of missingness separately. We do not recommend using either the total  $DIC_C$  or the model of interest part of  $DIC_C$ . However, we propose that the part of  $DIC_C$  relating to the model of missingness,  $f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$ , can be used for comparing the fit of the model of missingness for selection models with the same model of interest.

## 2.4 DIC based on the observed data likelihood

An alternative construction is based on the observed data likelihood,  $L(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m})$ ,

$$DIC_O = 2 \log L\{E(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}) | \mathbf{y}_{obs}, \mathbf{m}\} - 4 E_{\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}} \{\log L(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m})\}$$

where

$$L(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}) \propto \int f(\mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m} | \boldsymbol{\lambda}, \boldsymbol{\theta}) d\mathbf{y}_{mis}.$$

For a selection model, recalling Equation 1:

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m}) &\propto \int f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}) f(\mathbf{y}_{obs}, \mathbf{y}_{mis} | \boldsymbol{\lambda}) d\mathbf{y}_{mis} \\ &= f(\mathbf{y}_{obs} | \boldsymbol{\lambda}) E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\lambda}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})\}. \end{aligned} \quad (5)$$

$DIC_O$  differs from  $DIC_C$  in the model of missingness part, which is evaluated by integrating over  $\mathbf{y}_{mis}$  rather than by conditioning on it. The calculation of the expectation in Equation 5 creates complexity in the  $DIC_O$  computation.

The fit of the model of interest to  $\mathbf{y}_{obs}$  is optimised if this part of the model is estimated in isolation, i.e. we assume ignorable missingness. As soon as we allow for informative missingness by estimating the model of interest jointly with the model of missingness, the  $\boldsymbol{\lambda}$  estimates are influenced by the model of missingness. Hence if there is informative missingness, then the fit of the model of interest part to  $\mathbf{y}_{obs}$  necessarily deteriorates because in a selection model the same model of interest is applied to both the observed and the missing data.

$DIC_O$  is based on the joint model, with the model of interest and model of missingness contributing equally to differences in  $DIC_O$  between two joint models. As the model of interest contribution is based only on the observed data, the value of the part of the  $DIC_O$  attributable to the fit of the model of interest will increase the further the model of missingness departs from MAR. This can lead to a higher  $DIC_O$  if not offset by improvements in the fit of the model of missingness part (as happens in our example in Section 5), and so it should not be used to select a model of missingness. In order to

make comparisons using  $DIC_O$ , we need to fix the model of missingness part and use it to help choose the model of interest. Even so, we must still be careful how we use  $DIC_O$ , remembering that it only tells us about the fit of a selection model to the observed data and nothing about its fit to the missing data.

## 2.5 Plug-ins for calculating DIC

In general, the required plug-in likelihood can be calculated by using plug-ins on different scales. In a regression framework these can be the stochastic parameters used to calculate the linear predictor or the linear predictor itself. This choice affects both parts of the model, but to illustrate we consider the model of missingness. Let the model of missingness be defined as

$$\begin{aligned} m_i &\sim \text{Bernoulli}(p_i), \\ \text{link}(p_i) &= f(y_i, \boldsymbol{\theta}), \\ \boldsymbol{\theta} &\sim \text{prior distribution.} \end{aligned} \tag{6}$$

where ‘link’ would typically be taken to be the logit or probit function. Then, we could use the  $\boldsymbol{\theta}$  as our plug-in values ( $\boldsymbol{\theta}$  plug-ins) or alternatively we might use the  $\text{link}(p_i)$  as plug-ins (**linkp** plug-ins). WinBUGS uses the former, taking the posterior means  $\hat{\theta}_k = E(\theta_k)$  for all  $\theta_k$  in the set of model of missingness parameters as the plug-ins. If we use the **linkp** plug-ins to calculate  $DIC_C$  instead, then our plug-ins are evaluated as  $E(\boldsymbol{\lambda}, \text{linkp} | \mathbf{y}_{obs}, \mathbf{m})$  rather than  $E(\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})$ . These plug-ins lead to a different plug-in deviance, and hence a different DIC which may affect our model comparison.

For  $DIC_O$  we must choose plug-ins that ensure consistency in the calculation of the posterior mean deviance and the plug-in deviance, so that missing values are integrated out in both parts of the DIC. The  $\boldsymbol{\theta}$  plug-ins allow us to evaluate  $\hat{D}$  by integrating over  $\mathbf{y}_{mis}$  as required. By contrast, the **linkp** plug-ins are not appropriate as they do not allow averaging over a sample of  $\mathbf{y}_{mis}$  values, and in fact would lead to the same plug-in deviance as the one used for calculating  $DIC_C$ . Hence for  $DIC_O$ , we only use  $\boldsymbol{\theta}$  plug-ins, and the  $DIC_O$  plug-ins must be evaluated as  $E(\boldsymbol{\lambda}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{m})$ .

## 2.6 Strategy for using DIC to compare selection models

Suppose that we have a number of models of interest that are plausible for the question under investigation, and a number of models of missingness that are plausible for describing the missingness mechanism that generated the missing outcomes. Then our proposed strategy is to fit a set of joint models, that combine each model of interest with each model of missingness.  $DIC_O$  can then be used to compare models with the same model of missingness, and the model of missingness part of  $DIC_C$  can be used to compare models with the same model of interest. Hence, by combining complementary information provided by the two DIC measures we contend that we can usefully assess the comparative fit of a set of models, whereas this is not possible with a single DIC measure.

$DIC_O$  cannot be computed using WinBUGS alone, because in general the required expectations cannot be evaluated directly from the output of a standard MCMC run. For these, either “nested” MCMC is required, or some other simulation method. In the next section, we discuss the steps involved in calculating  $DIC_O$  for a selection model where  $f(\mathbf{y} | \boldsymbol{\beta}, \sigma)$  is the model of interest, typically a linear

regression model assuming Normal or t errors in our applications, and  $f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta})$  is a commonly used Bernoulli model of non-ignorable missingness.

### 3 Implementation of $DIC_O$

Daniels and Hogan (2008) (henceforth DH) provide a broad outline of an algorithm for calculating  $DIC_O$ , which we use as a starting point for our implementation which uses the R software with calls to WinBUGS to carry out MCMC runs where necessary (the code will be made available on <http://www.bias-project.org.uk/>). We start by describing our preferred algorithm and then explain how and why it differs from the suggestions of DH. We then discuss issues connected with the choice of plug-ins, and the checks that we consider necessary to ensure that the samples generated for its calculation are of sufficient length.

#### 3.1 Algorithm

Our preferred algorithm, used to calculate the  $DIC_O$  for selection models implemented in the two examples in Sections 4 and 5, can be summarised by the following steps (fuller detail is provided in the Appendix):

- 1 Call WinBUGS to carry out a standard MCMC run on the selection model, and save samples of length  $K$  of the model of interest and model of missingness parameters, denoted  $\boldsymbol{\beta}^{(k)}$ ,  $\sigma^{(k)}$  and  $\boldsymbol{\theta}^{(k)}$ ,  $k = 1, \dots, K$ , ( $Ksample$ ).
- 2 Evaluate the  $Ksample$  posterior means of the model parameters,  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\sigma}$  and  $\hat{\boldsymbol{\theta}}$ .
- 3 For each member,  $k$ , of the  $Ksample$ , generate a sample of length  $Q$  of missing responses from the appropriate likelihood evaluated at  $\boldsymbol{\beta}^{(k)}$  and  $\sigma^{(k)}$  using  $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\boldsymbol{\beta}^{(k)}, \sigma^{(k)})$  (the sample associated with member  $k$  of the  $Ksample$  is the  $Qsample^{(k)}$ ).
- 4 Next, for each member,  $k$ , of the  $Ksample$ , evaluate the expectation term from Equation 5,  $E_{\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}}\{f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\}$ , by averaging over its associated  $Qsample$ . Using these expectations, calculate the posterior mean of the deviance,  $\bar{D}$ , by averaging over the  $Ksample$ . (See step 4 in the Appendix for the required equations.)
- 5 Generate a new  $Qsample$  of missing responses from the appropriate likelihood evaluated at the posterior means of the model of interest parameters using  $f(\mathbf{y}_{obs}, \mathbf{y}_{mis}|\hat{\boldsymbol{\beta}}, \hat{\sigma})$ . Evaluate the expectation term of the plug-in deviance by averaging over this new  $Qsample$ , and calculate the plug-in deviance,  $\hat{D}$ , using the posterior means from the  $Ksample$ . (See step 5 in the Appendix for the required equations.)
- 6 Finally, calculate  $DIC_O = 2\bar{D} - \hat{D}$ .

The main differences between this algorithm and the DH proposal is in steps 3 and 4. DH propose using reweighting to avoid repeatedly generating samples for the evaluation of the expectations required in step 4. An implementation using reweighting involves generating a single  $Qsample$  of missing responses

from the appropriate likelihood evaluated at the posterior means of the model of interest parameters (as in step 5) instead of the multiple Qsamples at step 3. Step 4 then involves calculating a set of weights for each member of the Ksample, and using these in the evaluation of the expectation term. Fuller detail of the changes to these steps is provided in the Appendix.

The reweighting is a form of importance sampling, used when we wish to make inference about a distribution  $f^*(\cdot)$  using Monte Carlo integration, but instead have available a sample,  $z^{(1)}, \dots, z^{(Q)}$ , from a different distribution  $f(\cdot)$ . The available sample can be reweighted to make some inference based on  $f^*(\cdot)$ , using weights of the form

$$w_q = \frac{f^*(z^{(q)})}{f(z^{(q)})}.$$

Details of the equations for the weights required for calculating  $DIC_O$  using the reweighting method are given in the Appendix. The success of importance sampling is known to be critically dependent on the variability of the sampling weights (Peruggia, 1997), with greater variability leading to poorer estimates. For the method to be successful, we require that the two distributions  $f(\cdot)$  and  $f^*(\cdot)$  are reasonably close, and in particular that  $f(\cdot)$  has a heavier tail than  $f^*(\cdot)$  (Liu, 2001; Gelman *et al.*, 2004).

We have run both versions of the algorithm (with and without weighting) on some examples and recommend the version without weighting because it (1) avoids effective sample size problems associated with reweighting, (2) reduces instability and (3) has no computational disadvantage. We now discuss each of these issues in more detail.

### Effective sample size

In the calculation of  $DIC_O$  using reweighting, a set of sampling weights,  $w_q^{(k)}$ , is produced for each member of the Ksample. We would like the effective sample size (ESS) of each of these sets of weights to be close to the actual sample size,  $Q$ . Following Liu (2001), Chapter 2, we define ESS as

$$ESS = \frac{Q}{1 + var(w)} \tag{7}$$

where  $Q$  is the number of samples generated from a distribution,  $f(\cdot)$ , and  $var(w)$  is the variance of normalised importance weights. A small variance is good as ESS approaches the actual sample size,  $Q$ , when  $var(w)$  gets smaller. This variance can be estimated by

$$\frac{\sum_{q=1}^Q (w_q - \bar{w})^2}{(Q - 1)\bar{w}^2} \tag{8}$$

where  $\bar{w}$  is the mean of the sample of  $w_q$ s. Using these formulae, a set of ESS values can be calculated, one corresponding to each Ksample member. We found that the ESS is highly variable in examples based on simulated data, including a sizeable proportion which are sufficiently low to be of concern. By using an algorithm without reweighting, we avoid potential problems associated with low ESS.

### Stability

We would like the calculated value of  $DIC_O$  to be stable, and not depend on the random number seed used to generate either the Ksample or Qsample. For an example based on data with simulated

missingness, we calculated  $DIC_O$  using the reweighted algorithm with  $K = 2,000$  and  $Q = 2,000$ . Firstly, we repeated the calculation four times, using different random number seeds for generating the Ksample, but the same random number seed to generate the Qsample. The variation between the  $DIC_O$  from the five calculations (original and four repetitions) was small (less than 1). Note that in this case although the Qsample is generated from the same random number seed, it will also differ between runs due to the Ksample differences. Secondly, we repeated the calculation another four times, but using the same random number seed to generate the Ksample and four different random number seeds for generating the Qsample. As the Ksample is generated from the same random number seed, any differences are attributable to variation in the Qsample. Both  $\bar{D}$  and  $\hat{D}$  now exhibit much larger variation, resulting in a difference between the highest and lowest  $DIC_O$  of about 6 which is sufficiently large to be a concern, given that rules of thumb suggest that differences of 3-7 in DIC should be regarded as important (Spiegelhalter *et al.*, 2002). (These results are shown in Table 8 of the Supplementary Material.) Repeating the exercise with  $Q$  increased to 10,000 lowered the variation only slightly.

Using the algorithm without reweighting resulted in much greater stability of  $\bar{D}$ , but  $\hat{D}$ , and hence  $DIC_O$ , remained variable. A method for assessing this instability is discussed in Section 3.3.

### Computational time

One of the original reasons for using reweighting was to speed up the computation of  $DIC_O$ , since our preferred method involves generating  $K + (K \times Q) + Q$  samples, whereas the importance sampling method just generates  $K + Q$  samples, and then reweights the single  $Q$  sample for every replicate in the  $K$  sample. However, for equivalent sample sizes we found that our implementation of both algorithms ran in about the same time, so there appears to be no computational advantage to using reweighting in practice. This is because the computational time saved in the reweighting algorithm by not generating the extra samples, is offset by evaluating the weights which also requires the calculation of  $K \times Q$  model of interest likelihoods.

### 3.2 Choice of plug-in

As with calculating any DIC using posterior mean plug-ins, checks that these posterior means come from approximately symmetric unimodal distributions are essential. One possibility is a visual inspection of the posterior distributions of the proposed plug-ins and a check that the coefficient of skewness, where

$$\text{coefficient of skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\text{sd}(x)^3}, \quad (9)$$

which is a measure of the asymmetry of a distribution, is close to 0.

### 3.3 Adequacy of the size of the Qsample

As discussed above (see paragraph headed ‘‘Stability’’ in Section 3.1), we would like to be sure that  $Q$  is large enough to ensure that the  $DIC_O$  resulting from our calculations is stable. We have developed a method for checking the stability of our results using subsets of the Qsample. These subsets are created

by splitting the complete  $Q_{\text{sample}}$  in half, and then successively splitting the resulting subsets.  $\bar{D}$  and  $\hat{D}$  for each subset and the full sample are then plotted against the size of the  $Q_{\text{sample}}$ . ( $\text{DIC}_O$  could also be plotted against  $Q_{\text{sample}}$  size, but as it is a function of  $\bar{D}$  and  $\hat{D}$ , it provides no additional information.) The required extra calculations can be carried out with negligible additional cost in running time.

Figure 1 provides examples of such plots, where  $Q = 40,000$  and the sample is repeatedly split until a sample size of 2,500 is reached. This gives 2 non-overlapping  $Q_{\text{samples}}$  of length 20,000, 4 non-overlapping  $Q_{\text{samples}}$  of length 10,000, 8 non-overlapping  $Q_{\text{samples}}$  of length 5,000 and 16 non-overlapping  $Q_{\text{samples}}$  of length 2,500. These plots show little variation in  $\bar{D}$  at each  $Q$  (all the crosses are on top of each other), but a clear downwards trend as  $Q$  increases, which converges towards a limit. However,  $\hat{D}$  exhibits instability for the plots labelled JM1 and especially for JM3, that decreases as  $Q$  increases. (We will explain the stability in the JM2 plot in Section 4.) A similar downwards trend to  $\bar{D}$ , converging to a limit is indicated by the mean values of  $\hat{D}$ .

We consider the  $Q_{\text{sample}}$  size to be sufficient if our proposed deviance plot suggests both  $\bar{D}$  and  $\hat{D}$  have converged to a limit and  $\hat{D}$  has stabilised. On this basis 40,000 appears an adequate sample size for calculating  $\text{DIC}_O$  for JM1, but a higher  $Q$  might produce a more accurate  $\text{DIC}_O$  for JM2 and JM3. The plots for this and other synthetic examples suggest that higher variability and slower convergence to a limit are associated with poorer fitting models.

## 4 Illustration of strategy on simulated data

We now assess how  $\text{DIC}_O$  and the missingness part of  $\text{DIC}_C$  can be used to help choose between models, using simulated data with simulated missingness so that the correct model is known. For this purpose, we generate a dataset of bivariate Normal data with 1000 records comprising a response,  $y$ , and a single covariate,  $x$ , s.t.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right). \quad (10)$$

For this dataset the correct model of interest is

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned} \quad (11)$$

and the true values of the parameters are  $\beta_0 = 1$  and  $\beta_1 = 0.5$ .

We then delete some of the responses according to the equation  $p_i = \phi_0 + \phi_1 y_i$ . The values of  $\phi_0$  and  $\phi_1$  are chosen to impose linear non-ignorable missingness with a steep positive gradient, such that the probability of being missing for the lowest value of  $y$  is 0 and the probability of being missing for the highest value of  $y$  is 1. The chosen values also ensure that  $0 \leq p_i \leq 1$  for all  $y_i$ . Although the true model of missingness is the linear equation  $p_i = \phi_0 + \phi_1 y_i$ , this can be adequately modelled by the linear logistic equation  $\text{logit}(p_i) = \theta_0 + \theta_1 y_i$ , which ensures that the estimated probabilities always lie in the range  $[0,1]$ .

Our investigation is based on fitting four joint models (JM1-JM4), as specified in Table 1, to this simulated dataset with simulated missingness. For JM1, we know that the specified model of interest

is correct and that the specified linear logit is a good approximation for the true model of missingness. However, JM2 has an inadequate model of interest, JM3 has an incorrect error distribution and JM4 has too complex a model of missingness. So for the simulated data, we consider three different models of interest and two different models of missingness. A full implementation of our proposed strategy would involve fitting a set of joint models which pairs each model of interest with each model of missingness (six joint models), and we do this in our real data application in Section 5.

Table 1: Specification of joint models for the simulated data

Model Name	Model of Interest	Model of Missingness
JM1	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM2	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM3	$y_i \sim t_4(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM4	$y_i \sim N(\mu_i, \sigma^2); \quad \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i + \theta_2 y_i^2$

Vague priors are specified for the unknown parameters of the model of interest: the  $\beta$  parameters are assigned  $N(0, 10000^2)$  priors and the precision,  $\tau = \frac{1}{\sigma^2}$ , a  $\text{Gamma}(0.001, 0.001)$  prior. Following Wakefield (2004) and Jackson *et al.* (2006), we specify a  $\text{logistic}(0, 1)$  prior for  $\theta_0$  and a weakly informative  $N(0, 0.68)$  prior for  $\theta_1$  and  $\theta_2$ , which corresponds to an approximately flat prior on the scale of  $p_i$ .

We calculate the  $\text{DIC}_O$  for the three models with the same model of missingness (JM1, JM2 and JM3) using the algorithm described in Section 3, with  $K = 2,000$  and  $Q = 40,000$ . The likelihood for the model of interest is calculated using

$$f(\mathbf{y}|\boldsymbol{\beta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu_i)^2\right) \quad \text{for JM1\&JM2}$$

$$\text{and } f(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^n \frac{\Gamma(\frac{5}{2})}{2\sigma\sqrt{\pi}} \left[1 + \left(\frac{y_i - \mu_i}{2\sigma}\right)^2\right]^{-\frac{5}{2}} \quad \text{for JM3.}$$

In the  $t_4$  distribution  $\sigma$  is a scale parameter, s.t.  $\sigma = \sqrt{\frac{\text{var}}{2}}$ , where  $\text{var}$  is the variance of the distribution. For all three models, the likelihood for the model of missingness is calculated using

$$f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^n p_i^{m_i} (1 - p_i)^{1 - m_i}$$

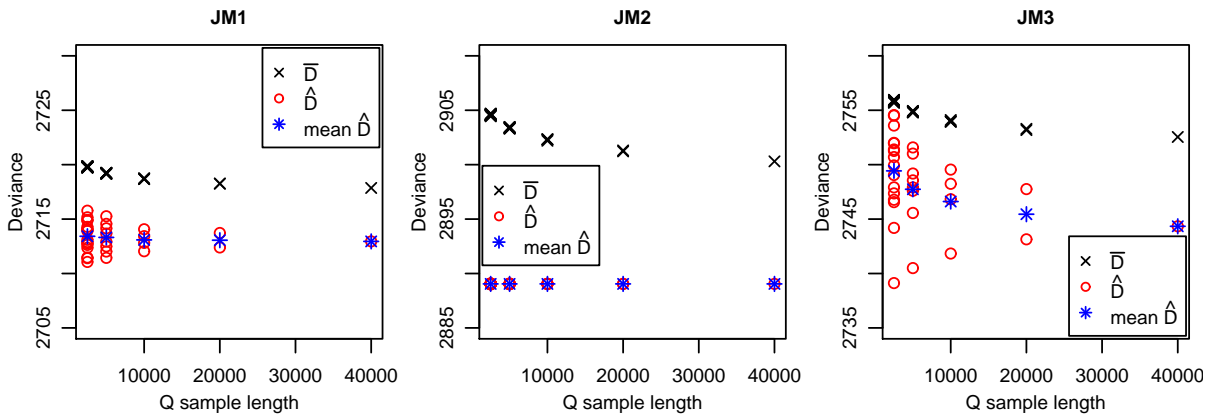
$$\text{where } p_i = \frac{e^{\theta_0 + \theta_1 y_i}}{1 + e^{\theta_0 + \theta_1 y_i}}.$$

The samples produced by the WinBUGS runs are from 2 chains of 15,000 iterations, with 10,000 burn-in and the thinning parameter set to 5. Based on the Gelman-Rubin convergence statistic (Brooks and Gelman, 1998) and a visual inspection of the chains, the WinBUGS runs required for calculating  $\text{DIC}_O$  for the three models all converged.

As recommended in Section 3, we start by checking that the posterior means used as plug-ins come from approximately symmetric unimodal distributions. From their coefficients of skewness (Equation 9) we find that the  $\boldsymbol{\theta}$  are skewed to some extent in all the models, and that the  $\mathbf{y}_{\text{mis}}$  are badly skewed for JM3 (see Table 9 in the Supplementary Material for figures). However, as this simple example is for illustration, we continue with these plug-ins, interpreting the  $\text{DIC}_O$  with caution.

As discussed in Section 3.3, Figure 1 allows us to assess the adequacy of the length of our  $Q_{\text{sample}}$  for the different models by splitting it into subsets and plotting  $\bar{D}$  and  $\hat{D}$  against the sample lengths. The scale ranges of the three plots are consistent, but the magnitudes vary. Downward trends in the deviance estimates, converging towards a limit, are shown for all models. However, for JM3 and possibly for  $\bar{D}$  in JM2, this limit appears not to have yet been reached. In this example, as  $\text{DIC}_O$  for these models is substantially higher than for JM1, we consider the sampling variability in  $\text{DIC}_O$  to be inconsequential by comparison and hence rerunning with larger  $Q$  to be unnecessary. Also, increased instability is evident in the  $\hat{D}$  estimates for JM3 compared to JM1. The stability in the  $\hat{D}$  for JM2 is because the posterior mean of  $\theta_1$  is close to 0, and so any variation in the  $\mathbf{y}_{\text{mis}}$  estimates between the  $Q_{\text{subsamples}}$  will have a minimal impact on the model of missingness likelihood and hence  $\hat{D}$ .

Figure 1: Deviance plots for checking the adequacy of the  $Q_{\text{sample}}$  length for JM1-JM3



Recall that since the data and missingness are simulated, we know that JM1 is the correct model. Table 2 shows the  $\text{DIC}_O$  for JM1-JM3, and an alternative measure of overall fit, the mean square error (MSE) for the model of interest, as defined by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - E(y_i|\boldsymbol{\beta}))^2, \quad (12)$$

where  $E(y_i|\boldsymbol{\beta})$  is evaluated as the posterior mean of  $\mu_i$  in Equation 11. The MSE is based only on the observed data, and slightly favours JM1, but there is little to choose between JM1 and JM3. JM2 is clearly a very poor choice. In contrast to the MSE,  $\text{DIC}_O$  assesses the joint fit of both parts of the model, penalised for complexity, although, as with MSE, the fit of the model of interest is only considered with respect to the observed responses.  $\text{DIC}_O$  suggests that JM1 is a better fitting model than JM2 or JM3.

Table 2: Comparison of  $\text{DIC}_O$  and MSE for JM1-JM3

	$\bar{D}$	$\hat{D}$	$p_D$	$\text{DIC}_O$	$\text{MSE}^a$
JM1	2717.9	2712.9	4.9	2722.8	0.762
JM2	2900.3	2889.0	11.3	2911.6	0.980
JM3	2752.5	2744.3	8.2	2760.8	0.774

<sup>a</sup> MSE based on observed data only

We now look at the model of missingness part of  $\text{DIC}_C$ , to help choose between JM1 and JM4, the two models with the same model of interest. In calculating  $\text{DIC}_C$  we are not restricted in our choice

of plug-ins for computational reasons as for  $DIC_O$ . So we calculate two versions, one automatically generated by WinBUGS using the  $\theta$  plug-ins and the other calculated using *linkp* plug-ins (now referred to as *logitp* plug-ins as we have specified a logistic link for this example). We find that the posterior distributions of the *logitp* plug-ins are very skewed for JM4 (Table 9 in the Supplementary material). However, the  $DIC_C$  based on  $\theta$  plug-ins is unusable for JM4 as it returns a negative  $p_D$ . A comparison of the model of missingness  $DIC_C$  based on the *logitp* plug-ins (Table 3) suggests clearly that the model of missingness in JM1 fits better than the JM4 model of missingness.

Table 3: Model of Missingness  $DIC_C$  for JM1 and JM4

	$\bar{D}$	<i>logitp</i> plug-ins		
		$\hat{D}$	$p_D$	$DIC_C$
JM1	1323.5	1297.9	25.6	1349.1
JM4	1305.5	1209.4	96.0	1401.5

Hence, if we use  $DIC_O$  to compare models with the same model of missingness, we will select JM1 in preference to JM2 and JM3, and if we use the model of missingness part of  $DIC_C$  to compare models with the same model of interest we will choose JM1 over JM4. So in this example, by combining information from these two DIC measures we successfully identify the correct model. We now examine this approach in a case study comparing three treatments of depression using longitudinal data.

## 5 Application

### 5.1 Description of HAMD data

As an application, we analyse data from a six centre clinical trial comparing three treatments of depression, which were previously analysed by Diggle and Kenward (1994) (DK) and Yun *et al.* (2007). In this clinical trial, 367 subjects were randomised to one of three treatments and rated on the Hamilton depression score (HAMD) on five weekly visits, the first before treatment, week 0, and the remaining four during treatment, weeks 1-4. The HAMD score is the sum of 16 test items and takes values between 0 and 50, where the higher the score the more severe the depression. In this example, we are interested in any differences between the effects of the three treatments on the change in depression score (HAMD) over time. Some subjects dropped out of the trial from week 2 onwards, with approximately one third lost by the end of the study. Similar numbers of subjects received each treatment (120, 118 and 129 for treatments 1, 2 and 3 respectively), but the levels and timing of drop-out differ. In particular, fewer subjects drop out of treatment 3, and although the missingness percentage is similar for treatments 1 and 2 by week 4, more of the drop-out occurs earlier for treatment 2. Hence there is a concern that non-ignorable missingness could have occurred linked to the treatment effect investigated. For the HAMD data, we wish to compare the fit of six joint models formed by combining two different models of interest with three different models of missingness, which we now describe.

## 5.2 Models of interest for HAMD data

Exploratory plots indicate a downwards trend in the HAMD score over time, so for our model of interest we follow DK and regress HAMD against time, allowing a quadratic relationship and a different intercept for each centre. We use two variants of this model: an autoregressive model and a random effects model. In the first (AR), we specify

$$\begin{aligned} y_{iw} &= \mu_{iw} + \delta_{iw} \\ \mu_{iw} &= \beta_{c(i)} + \eta_{t(i)}w + \xi_{t(i)}w^2 \end{aligned} \quad (13)$$

where  $i$ =individual,  $t$ =treatment (1,...,3),  $c$ =centre (1,...,6) and  $w$ =week (0,...,4).  $c(i)$  and  $t(i)$  denote the centre and treatment of individual  $i$  respectively. The  $\delta_{iw}$ s follow a second-order autoregressive process defined by

$$\begin{aligned} \delta_{i0} &= \epsilon_{i0}, \\ \delta_{i1} &= \alpha_1\delta_{i0} + \epsilon_{i1}, \\ \delta_{iw} &= \alpha_1\delta_{i(w-1)} + \alpha_2\delta_{i(w-2)} + \epsilon_{iw}, \quad w \geq 2 \\ \epsilon_{iw} &\sim N(0, \sigma^2). \end{aligned} \quad (14)$$

In the second (RE), we allow individual random effects on the intercept s.t.

$$\begin{aligned} y_{iw} &\sim N(\mu_{iw}, \sigma^2) \\ \mu_{iw} &= \beta_i + \eta_{t(i)}w + \xi_{t(i)}w^2 \\ \beta_i &\sim N(\gamma_{c(i)}, \rho_{c(i)}^2). \end{aligned} \quad (15)$$

For both variants we assign vague priors to the unknown parameters: giving the regression coefficients  $N(0,10000)$  priors and the precision ( $\frac{1}{\sigma^2}$ ) a Gamma(0.001,0.001) prior. In the RE version, each  $\gamma_{c(i)}$  is assigned a  $N(0,10000)$  prior and the hierarchical standard deviations  $\rho_{c(i)}$  are assigned noninformative uniform priors (with an upper limit of 100) as suggested by Gelman (2005).

## 5.3 Models of missingness for HAMD data

We specify three models of missingness as detailed in Table 4, and assign a logistic prior to  $\theta_0$  and weakly informative Normal priors to all the other  $\theta$  parameters as previously discussed (Section 4). The simplest form of informative drop-out is given by MoM1 where missingness depends on the current value of the HAMD score, while the form of MoM2 allows dependence on the previous week's HAMD score and the change in the HAMD score as parameterised by  $\theta_3$ . MoM3 has the same form as MoM2, but includes separate  $\theta$  for each treatment, which allows treatment to directly affect the missingness process.

Table 4: Specification of the models of missingness for the HAMD data

Model Name	Model of Missingness equation
MoM1	$\text{logit}(p_{iw}) = \theta_0 + \theta_1 y_{iw}$
MoM2	$\text{logit}(p_{iw}) = \theta_0 + \theta_2 y_{i(w-1)} + \theta_3 (y_{iw} - y_{i(w-1)})$
MoM3	$\text{logit}(p_{iw}) = \theta_{0t(i)} + \theta_{2t(i)} y_{i(w-1)} + \theta_{3t(i)} (y_{iw} - y_{i(w-1)})$

## 5.4 Comparison of joint models for HAMD data

Joint models combining the model of missingness MoM1 with the RE and AR models of interest will be referred to as JM1(RE) and JM1(AR) respectively, and similarly for models of missingness MoM2 and MoM3. Runs of these six joint models and the models of interest estimated on complete cases only, CC(RE) and CC(AR), converged based on the Gelman-Rubin convergence statistic and a visual inspection of the chains. Adding a missingness model makes little difference to the  $\beta$  or  $\gamma$  estimates, but there are substantial changes in some of the  $\eta$  and  $\xi$  parameters associated with the effect of treatment over time. The impact of these changes will be assessed shortly using plots of the mean response profiles for each treatment.

The model of missingness parameter estimates are shown in Table 5. The positive  $\theta_1$  estimates for the JM1 models suggest that drop-out is associated with high HAMD scores, while the negative  $\theta_3$  in the JM2 models indicate that change in the HAMD score is informative, with individuals more likely to drop-out if their HAMD score goes down. These two complementary messages are that the more severely depressed subjects, and those for whom the treatment appears most successful are more likely to drop-out. The JM3 models provide some evidence that the missingness process is affected by the treatment. These findings hold for both the AR and RE models.

Table 5: Parameter estimates for model of missingness for HAMD example

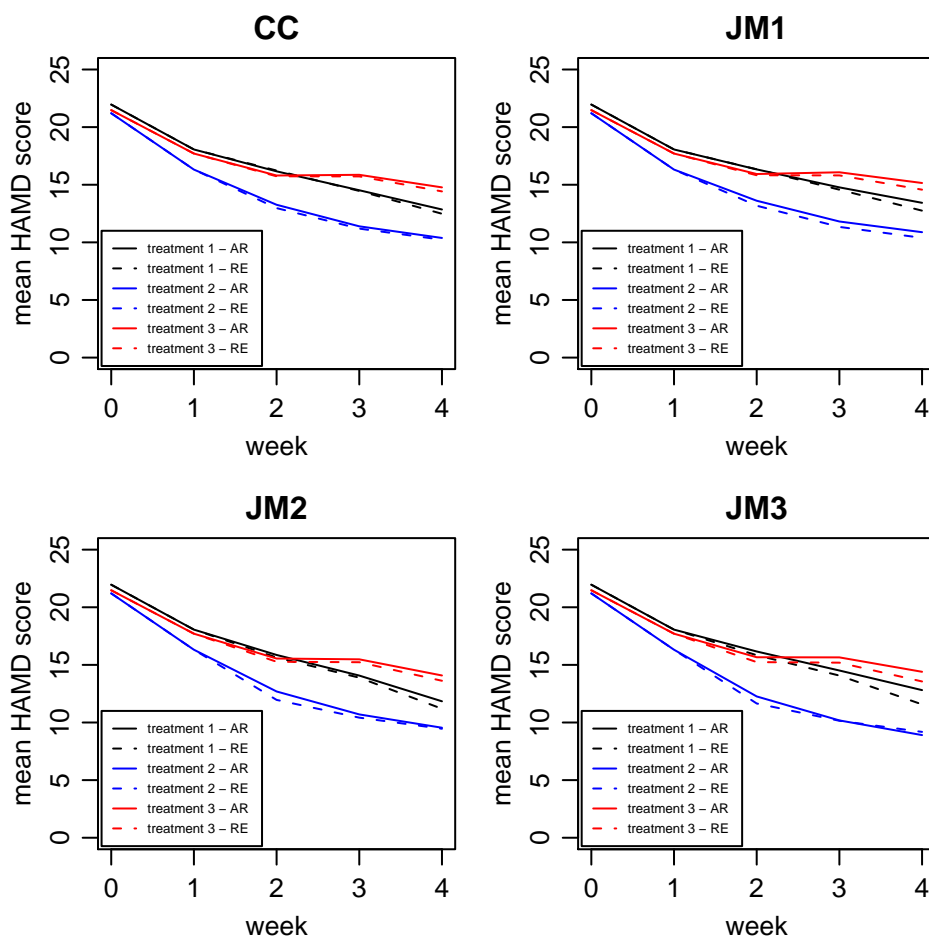
	JM1(AR)	JM2(AR)	JM3(AR)	JM1(RE)	JM2(RE)	JM3(RE)
$\theta_0$	-3.12 (-3.72,-2.53)	-3.19 (-3.80,-2.62)		-2.61 (-3.22,-2.03)	-3.10 (-3.75,-2.50)	
$\theta_0(t_1)$			-2.65 (-3.91,-1.58)			-2.22 (-3.22,-1.31)
$\theta_0(t_2)$			-3.75 (-5.20,-2.56)			-3.79 (-5.38,-2.41)
$\theta_0(t_3)$			-3.89 (-5.10,-2.81)			-3.57 (-4.87,-2.38)
$\theta_1$	0.08 (0.04,0.11)			0.04 (0.01,0.08)		
$\theta_2$		0.04 (0.00,0.09)			-0.01 (-0.07,0.04)	
$\theta_2(t_1)$			0.04 (-0.04,0.12)			-0.02 (-0.10,0.05)
$\theta_2(t_2)$			0.01 (-0.10,0.10)			-0.10 (-0.27,0.03)
$\theta_2(t_3)$			0.08 (0.00,0.15)			-0.01 (-0.12,0.09)
$\theta_3$		-0.14 (-0.27,-0.02)			-0.28 (-0.39,-0.18)	
$\theta_3(t_1)$			0.00 (-0.21,0.27)			-0.17 (-0.32,-0.04)
$\theta_3(t_2)$			-0.34 (-0.59,-0.10)			-0.54 (-0.87,-0.30)
$\theta_3(t_3)$			-0.08 (-0.28,0.08)			-0.32 (-0.54,-0.13)

Table shows the posterior mean, with the 95% interval in brackets

### 5.4.1 How much difference does the choice of model of interest make?

To see whether using AR or RE as our model of interest makes a difference, we compare the mean response profiles for each pair of models (Figure 2). For a complete case analysis the solid (AR) and dashed (RE) lines for each treatment are almost identical, and there are very small differences for JM1 and JM2, which accentuate for the more complex JM3. This is consistent with the discussion in Daniels and Hogan (2008) concerning the increased importance of correctly specifying the dependence structure in the model of interest when dealing with missing data.

Figure 2: Modelled mean response profiles for HAMD example - comparing the model of interest



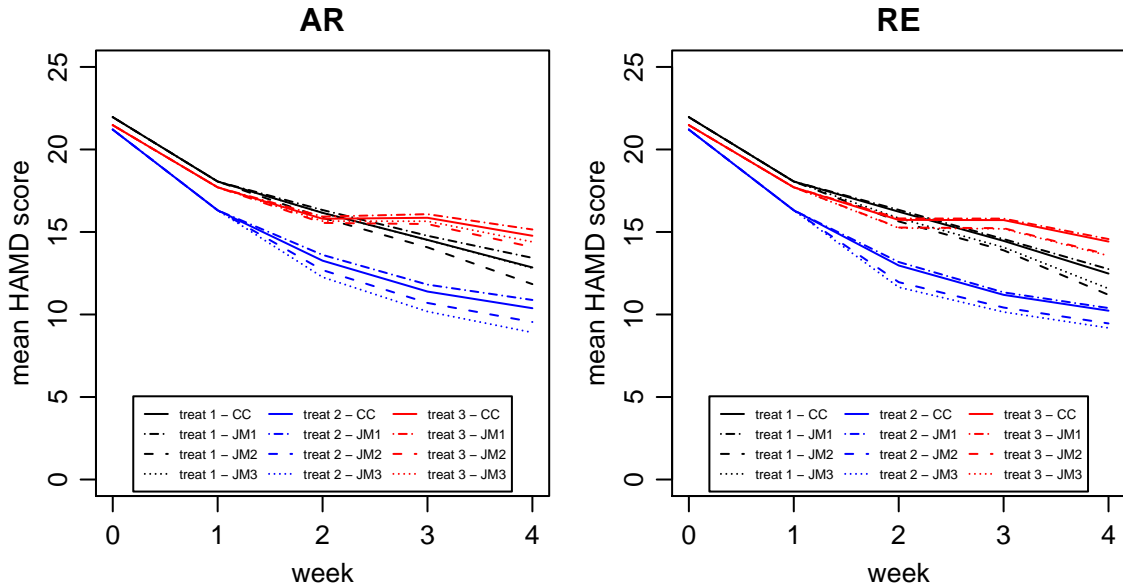
#### 5.4.2 How much difference does the choice of model of missingness make?

The impact of the addition of the MoM1 model of missingness to the AR model of interest can be seen by comparing the CC (solid lines) and JM1 (dot-dash lines) in Figure 3 and noticing a small upward shift of JM1; the impact is hardly discernible when the RE model of interest is used. By contrast, there is a consistent downwards shift from CC when MoM2 is added to both models of interest (dashed lines). However, adding MoM3 (dotted lines) shifts the CC profiles by different amounts resulting in increased differences between treatments, particularly for the AR model of interest.

#### 5.5 Use of $DIC_O$ to help with model choice

$DIC_O$  is calculated for the six HAMD models using the algorithm discussed in Section 3 and given more fully in the Appendix. The runs using MoM1 and MoM2 take approximately 5 hours on a desktop computer with a dual core 2.4GHz processor and 3.5GB of RAM, while the more complex models with MoM3 run in about 24 hours. As before we shall refer to the samples generated at steps 1 and 3 as our ‘Ksample’ and ‘Qsample’ respectively, with  $K$  and  $Q$  denoting the lengths of these samples. The Ksample is set to 2,000, formed from 2 chains of 110,000 iterations, with 100,000 burn-in and the thinning parameter set to 10, and  $Q$  is set to 40,000. Table 6 shows  $DIC_O$  for our six models

Figure 3: Modelled mean response profiles for HAMD example - comparing the model of missingness



In the AR plot, the CC and JM3 lines for treatment 1 are almost coincident. In the RE plot, the CC and JM1 lines are almost coincident for all treatments, and the JM2 and JM3 lines for treatment 3 are almost coincident.

for the HAMD data. The likelihood for the model of missingness is calculated for the weeks with drop-out, and for each of these weeks excludes individuals who have already dropped out.

Table 6:  $DIC_O$  for HAMD example

	$\bar{D}$	$\theta$ plug-ins		
		$\hat{D}$	$p_D$	$DIC_O$
JM1(AR)	9995.8	9978.6	17.2	10013.0
JM1(RE)	9663.2	9359.6	303.7	9966.9
JM2(AR)	9991.0	9965.5	25.5	10016.5
JM2(RE)	9680.6	9372.6	308.0	9988.5
JM3(AR)	9995.1	9965.0	30.1	10025.2
JM3(RE)	9698.1	9392.1	306.0	10004.2

Before discussing these results, we check that they are soundly based. Looking at the coefficients of skewness for the posterior distributions of the plug-ins (not shown), we find that some are skewed, most notably for the two JM3 models. To examine the adequacy of the  $Q$ sample, we split it into subsets and plot  $\bar{D}$  and  $\hat{D}$  against the sample lengths as described in Section 3. From these plots for the six models (shown as Figure 4 in the Supplementary Material), we see that both  $\bar{D}$  and  $\hat{D}$  are stable and show little variation even for small  $Q$  for both JM1 models. For the other models, trends similar to those exhibited by our synthetic data (see Figure 1) are evident, but again there is convergence to a limit suggesting the adequacy of  $Q=40,000$ . As before, we also see that the instability associated with small  $Q$  decreases with increased sample size. The trends and variation are more pronounced for the RE models than the AR models.

Our investigation with simulated data suggests that  $DIC_O$  can give useful information about which model of interest to select. For the HAMD example,  $DIC_O$  provides consistent evidence that the

random effects model of interest is preferable to the autoregressive model of interest when combined with each model of missingness, as can be seen by  $DIC_O$  always being smaller for RE than AR for each of the three models of missingness.

## 5.6 Use of the model of missingness $DIC_C$ to help with model choice

We now turn to the model of missingness part of  $DIC_C$ , to see what additional information it provides. Table 7 shows two versions, one based on the  $\theta$  plug-ins and one on the *logitp* plug-ins. As with the  $\theta$  plug-ins, the posterior distributions of the *logitp* plug-ins become increasingly skewed as the model of missingness becomes more complex. We have reservations about both sets of plug-ins, but find that they provide a consistent message from the model of missingness  $DIC_C$ . MoM2 and MoM3, used in the JM2 and JM3 models, provide clearly a better fit to this part of the model than JM1, with an edge towards JM3 rather than JM2, i.e. a missingness model that allows treatment specific parameters.

Table 7: Model of missingness  $DIC_C$  for HAMD example

	$\bar{D}$	$\theta$ plug-ins			<i>logitp</i> plug-ins		
		$\hat{D}$	$p_D$	$DIC_C$	$\hat{D}$	$p_D$	$DIC_C$
HAMD.JM1(AR)	698.6	695.5	3.1	701.8	694.3	4.3	703.0
HAMD.JM2(AR)	653.4	649.5	3.9	657.3	636.5	16.9	670.2
HAMD.JM3(AR)	626.0	621.8	4.2	630.2	583.1	42.9	668.9
HAMD.JM1(RE)	719.6	717.8	1.9	721.5	716.3	3.3	723.0
HAMD.JM2(RE)	547.5	517.7	29.8	577.3	511.2	36.3	583.8
HAMD.JM3(RE)	521.6	480.4	41.2	562.8	464.6	57.0	578.6

## 5.7 Combined use of $DIC_O$ and the model of missingness $DIC_C$

To conclude,  $DIC_O$  suggests that the RE model of interest is better than the AR. For RE models, there are substantial improvements in the model of missingness  $DIC_C$  for JM2 and JM3 over JM1, i.e. JM2 and JM3 better explain the missingness pattern than JM1. Overall, of the joint models explored, those with a RE model of interest and a model of missingness that depends on the change in HAMD (either treatment specific or not) seem most appropriate for the HAMD data.

If we based our analysis of this clinical trial data on a complete case analysis, we would conclude that treatment 2 lowers the HAMD score more than treatments 1 and 3 throughout the trial, and treatment 1 is more successful than treatment 3 in lowering HAMD in the later weeks. Using our preferred joint models, although all the treatments appear a little more effective in lowering HAMD, their ordering is unaltered (compare the dotted and dashed lines with the solid lines in the RE plot of Figure 3).

## 6 Discussion

For complete data, DIC is routinely used by Bayesian statisticians to compare models, a practice facilitated by its automatic generation in WinBUGS. However, using DIC in the presence of missing data is far from straightforward. The usual issues surrounding the choice of plug-ins are heightened, and in addition we must ensure that its construction is sensible. The conditional DIC automatically generated by WinBUGS, the total  $DIC_C$ , has asymptotic and coherency difficulties for selection models. We do not recommend using either the total  $DIC_C$  or the model of interest  $DIC_C$  or any single measure of DIC in isolation. The model choice strategy that we have developed relies on using both the model of missingness part of  $DIC_C$  and  $DIC_O$ . The model of missingness part of  $DIC_C$  allows comparison of the fit of the model of missingness for selection models with the same model of interest. A DIC based on the observed data likelihood,  $DIC_O$ , can help with the choice of the model of interest, and should be used to compare joint models built with the same model of missingness but different models of interest.

$DIC_O$  cannot be generated by WinBUGS, but can be calculated from WinBUGS output using other software. DH provide an algorithm for its calculation, which we have adapted and implemented for both simulated and real data examples. We recommend performing two sets of checks: (1) that the plug-ins are reasonable (i.e. if posterior means are used, they should come from symmetric, unimodal posterior distributions, and they must ensure consistency in the calculation of the posterior mean deviance and the plug-in deviance, so that missing values are integrated out in both parts of the DIC) and (2) that the size of the samples generated from the likelihoods ( $Q_{\text{sample}}$ ) is sufficiently large to avoid overestimating  $DIC_O$  and problems with instability in the plug-in deviance (we suggest plotting deviance against sample length and checking for stability, as in Figure 1). Based on limited exploration of synthetic and real data, we tentatively propose working with a  $Q_{\text{sample}}$  of at least 40,000. Again based on our experience, we tentatively suggest that even with a well chosen  $Q_{\text{sample}}$  size, a DIC difference of at least 5 is required to provide some evidence of a genuine difference in the fit of two models, as opposed to reflecting sampling variability.

In using  $DIC_O$  we must remember that it will only tell us about the fit of our model to the observed data and nothing about the fit to the missing data. However, it does seem reasonable to use it to compare joint models with different models of interest but the same models of missingness. DH discussed an alternative construction ( $DIC_F$ ) for selection models based on the posterior predictive expectation of the full data likelihood,  $L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \mathbf{m})$ , and provided a broad outline for its implementation.  $DIC_F$  may provide additional information for model choice, but its calculation is complicated as the expectation for the plug-ins is conditional on  $\mathbf{y}_{\text{mis}}$  and is beyond the scope of this paper. An alternative to using DIC to compare models, is to assess model fit using a set of data not used in the model estimation, if available.

Although the model of missingness  $DIC_C$  and  $DIC_O$  can provide complementary, useful insights into the comparative fit of various selection models, it would be a mistake to use them to select a single model. Even with straightforward data, such as our simulated example, the usual plug-ins are affected by skewness. This skewness makes the interpretation of DIC more complicated, as we have to allow for some additional variability that can obscure the message from the proposed strategy. Given this and the lack of knowledge regarding the fit of the missing data, DIC should never be used in isolation. We consider that sensitivity analysis to check that conclusions are robust to a range of assumptions

about the missing data remains crucial. However, our investigations have shown that these two DIC measures have the potential to assist in the selection of a range of plausible models, which allow the uncertainty introduced by non-ignorable missing data to be propagated into conclusions about a question of interest.

## Appendix

Our preferred algorithm for calculating  $DIC_O$  proceeds as follows: ( $f(\mathbf{y}|\boldsymbol{\beta}, \sigma)$  is the model of interest, typically Normal or t in our applications, and  $f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta})$  is a Bernoulli model of missingness in a selection model)

- 1 Carry out a standard MCMC run on the joint model  $f(\mathbf{y}, \mathbf{m}|\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ . Save samples of  $\boldsymbol{\beta}$ ,  $\sigma$  and  $\boldsymbol{\theta}$ , denoted by  $\boldsymbol{\beta}^{(k)}$ ,  $\sigma^{(k)}$  and  $\boldsymbol{\theta}^{(k)}$ ,  $k = 1, \dots, K$ , which we shall call the *Ksample*.
- 2 Evaluate the posterior means of  $\boldsymbol{\beta}$ ,  $\sigma$  and  $\boldsymbol{\theta}$ , denoted by  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\sigma}$  and  $\hat{\boldsymbol{\theta}}$ . (Evaluate  $\hat{\sigma}$  on the log scale and then back transform, see discussion headed ‘‘Skewness in the plug-ins for the simulated example’’ in the Supplementary Material for rationale.)
- 3 For each member of the Ksample, generate a sample  $\mathbf{y}_{mis}^{(kq)}$ ,  $q = 1, \dots, Q$ , from the appropriate likelihood evaluated at  $\boldsymbol{\beta}^{(k)}$  and  $\sigma^{(k)}$ , e.g.  $y_{mis}^k \sim N(\mathbf{X}\boldsymbol{\beta}^{(k)}, \sigma^{(k)2})$ . We denote the sample associated with member  $k$  of the Ksample as  $Qsample^{(k)}$ .
- 4 Then evaluate

$$h^{(k)} = E_{\mathbf{y}_{mis}|\mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}} \{f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\} \approx \frac{1}{Q} \sum_{q=1}^Q f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(kq)}, \boldsymbol{\theta}^{(k)}).$$

Calculate the posterior expectation of the observed data log likelihood as

$$\log L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{m}) \approx \frac{1}{K} \sum_{k=1}^K \left[ \log L(\boldsymbol{\beta}^{(k)}, \sigma^{(k)}|\mathbf{y}_{obs}) + \log h^{(k)} \right].$$

Multiply this by -2 to get the posterior mean of the deviance, denoted  $\bar{D}$ .

- 5 Generate a new Qsample,  $\mathbf{y}_{mis}^{(q)}$ ,  $q = 1, \dots, Q$ , using  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}$ . Evaluate the plug-in observed data log likelihood using the posterior means from the Ksample as

$$\log L(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}|\mathbf{y}_{obs}, \mathbf{m}) \approx \log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}|\mathbf{y}_{obs}) + \log \left( E_{\mathbf{y}_{mis}|\mathbf{y}_{obs}, \hat{\boldsymbol{\beta}}, \hat{\sigma}} \{f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \hat{\boldsymbol{\theta}})\} \right)$$

where

$$E_{\mathbf{y}_{mis}|\mathbf{y}_{obs}, \hat{\boldsymbol{\beta}}, \hat{\sigma}} \{f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \hat{\boldsymbol{\theta}})\} \approx \frac{1}{Q} \sum_{q=1}^Q f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}^{(q)}, \hat{\boldsymbol{\theta}}).$$

Multiply this plug-in log likelihood by -2 to get the plug-in deviance, denoted  $\hat{D}$ .

- 6 Finally, calculate  $DIC_O = 2\bar{D} - \hat{D}$ .

To implement an algorithm using reweighting as proposed by DH, alter steps 3-5 as follows:

- 3 Generate a Qsample  $\mathbf{y}_{mis}^{(q)}$ ,  $q = 1, \dots, Q$ , from the appropriate likelihood evaluated at the posterior means, e.g.  $y_{mis} \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$  (as in step 5 of our preferred algorithm).

- 4 For each value of  $(\boldsymbol{\beta}^{(k)}, \sigma^{(k)})$  in the Ksample, and each value of  $\mathbf{y}_{mis}^{(q)}$  from the Qsample, calculate the weight

$$w_q^{(k)} = \frac{f(\mathbf{y}_{mis}^{(q)} | \mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)})}{f(\mathbf{y}_{mis}^{(q)} | \mathbf{y}_{obs}, \hat{\boldsymbol{\beta}}, \hat{\sigma})}.$$

and evaluate

$$h^{(k)} = E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\} \approx \frac{\sum_{q=1}^Q w_q^{(k)} f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}^{(q)}, \boldsymbol{\theta}^{(k)})}{\sum_{q=1}^Q w_q^{(k)}}.$$

Calculate the posterior expectation of the observed data log likelihood and  $\bar{D}$  as before.

- 5 There is no need to generate a further Qsample, simply use the Qsample generated at the replacement step 3 to evaluate the plug-in observed data log likelihood and  $\hat{D}$  as before.

### Acknowledgements

Financial support: this work was supported by an ESRC PhD studentship (Alexina Mason). Sylvia Richardson and Nicky Best would like to acknowledge support from ESRC: RES-576-25-0015. The authors are grateful to Mike Kenward for useful discussions and providing the clinical trial data analysed in this paper. Alexina Mason thanks Ian Plewis for his encouragement during her research on non-ignorable missing data.

## Supplementary Material

### Stability of $DIC_O$ calculations

The results of the repeated  $DIC_O$  calculations described in the paragraph headed “Stability” in Section 3.1 are shown in Table 8.

Table 8: Variability in  $DIC_O$  calculated by the reweighted algorithm due to using a different random number seed to generate the Ksample or Qsample,  $K=Q=2000$

	$\bar{D}$	$\hat{D}$	$p_D$	$DIC_O$
Original	2418.5	2411.0	7.4	2425.9
Repetition1a - Ksample seed changed <sup>a</sup>	2418.1	2410.8	7.3	2425.4
Repetition2a - Ksample seed changed <sup>a</sup>	2418.2	2411.0	7.2	2425.4
Repetition3a - Ksample seed changed <sup>a</sup>	2418.1	2410.9	7.2	2425.3
Repetition4a - Ksample seed changed <sup>a</sup>	2418.3	2411.0	7.4	2425.7
Repetition1b - Qsample seed changed <sup>b</sup>	2419.1	2410.4	8.8	2427.9
Repetition2b - Qsample seed changed <sup>b</sup>	2420.0	2413.5	6.4	2426.4
Repetition3b - Qsample seed changed <sup>b</sup>	2423.0	2416.0	7.0	2430.0
Repetition4b - Qsample seed changed <sup>b</sup>	2424.6	2417.6	7.0	2431.7

In this example, joint model JM1, as described in Table 1, has been repeatedly fitted to a subset of real test score data taken from the National Child Development Study, with simulated non-ignorable linear missingness.

<sup>a</sup> Each repetition uses a different random number seed to generate the Ksample, but the same random number seed to generate the Qsample.

<sup>b</sup> Each repetition uses the same random number seed to generate the Ksample, but a different random number seed to generate the Qsample.

## Skewness in the plug-ins for the simulated example

The coefficient of skewness of the posterior distribution for various plug-ins used in calculating  $DIC_O$  and the model of missingness part of  $DIC_C$  for the simulated example are shown in Table 9. Mean and 95% interval values are given for  $y_{mis}$ , and the *logitp* for all individuals, observed individuals and missing individuals. As a guide to interpreting the values for our Ksample of size 2,000, 95% of 10,000 simulated Normal datasets with 2,000 members had skewness in the interval (-0.1,0.1). Even in this straightforward simulated example, the usual plug-ins are affected by skewness, sometimes badly.  $\sigma$ ,  $\log(\sigma)$  and  $\tau$  are all included in the table, and the difference in their skewness demonstrates sensitivity to the choice of the form of the scale parameter plug-in. This provides evidence that using a log transformation for  $\sigma$  is appropriate as argued by Spiegelhalter *et al.* (2002). (All our DIC calculations work with plug-in values for  $\sigma$  calculated on the log scale.)

Table 9: Skewness of posterior distribution of plug-ins for simulated data (skewness outside the interval (-1,1) highlighted in bold)

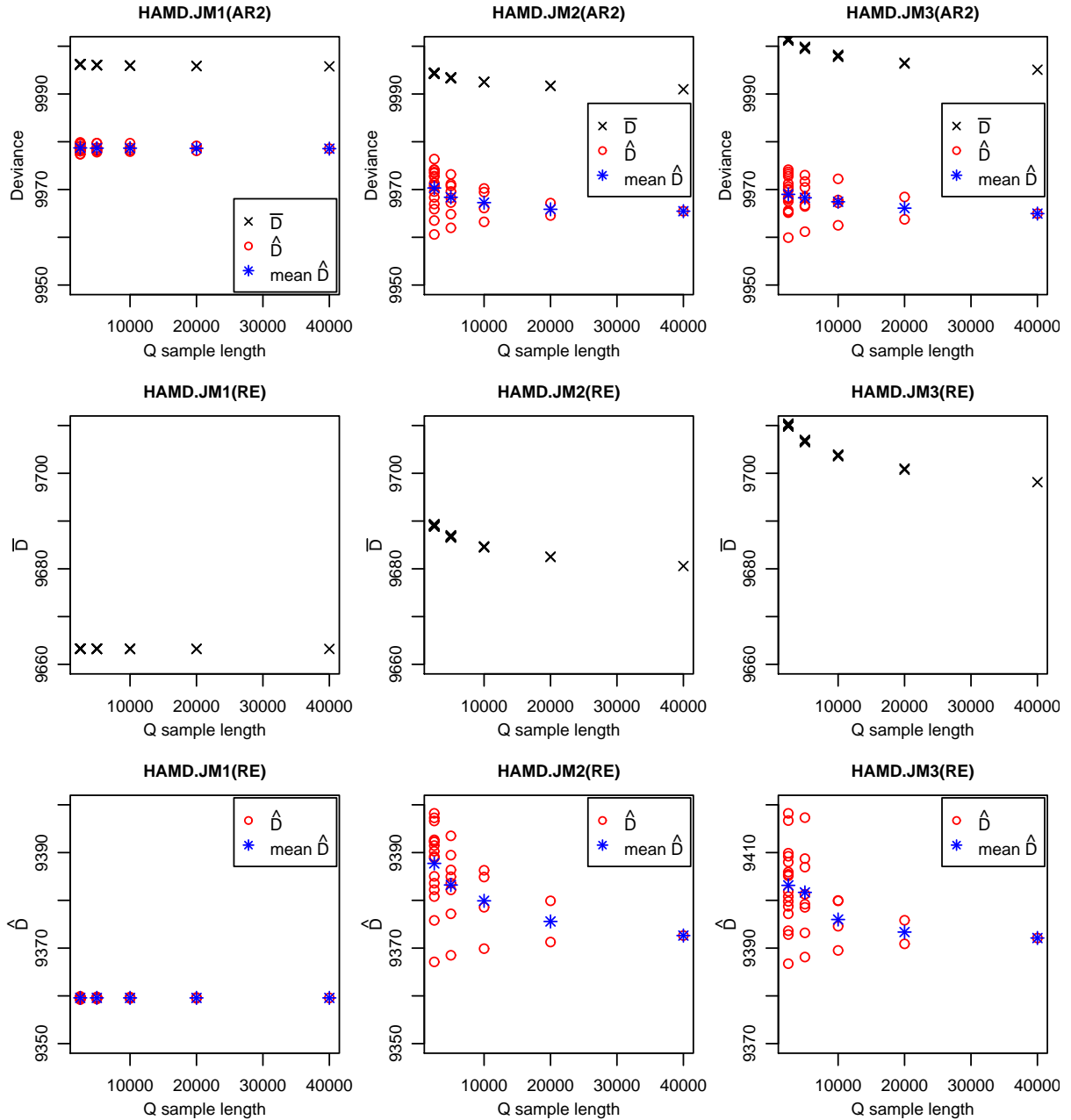
	JM1	JM2	JM3	JM4
$\beta_0$	0.006	0.079	0.098	0.198
$\beta_1$	-0.024		0.081	0.200
$\sigma$	0.187	<b>1.164</b>	0.194	0.360
$\log(\sigma)$	0.089	0.965	0.072	0.195
$\tau$	0.108	-0.604	0.175	0.126
$\theta_0$	-0.245	<b>-1.464</b>	-0.261	-0.325
$\theta_1$	0.107	0.089	0.120	0.415
$\theta_2$				-0.294
$y_{mis}^a$	0.005	0.016	<b>1.276</b>	-0.074
	(-0.088,0.094)	(-0.072,0.097)	(0.512,2.572)	(-0.196,0.045)
<i>logitp</i> (all) <sup>a</sup>	0.012	-0.168	0.647	<b>1.627</b>
	(-0.252,0.436)	(-1.818,1.451)	(-0.279,2.324)	(-0.636,5.374)
<i>logitp</i> (obs) <sup>a</sup>	-0.150	<b>-1.375</b>	-0.165	-0.311
	(-0.253,-0.012)	(-1.859,-0.665)	(-0.282,-0.013)	(-0.641,-0.087)
<i>logitp</i> (mis) <sup>a</sup>	0.195	<b>1.198</b>	<b>1.568</b>	<b>3.821</b>
	(-0.124,0.495)	(0.869,1.555)	(0.718,2.839)	(1.256,6.291)

<sup>a</sup> mean value is shown, with 95% interval below in brackets

## Checking the adequacy of the Qsample length for the HAMD example

Plots of  $\bar{D}$  and  $\hat{D}$  against the Qsample lengths for the six models in the HAMD example, as described in Section 5.5, are displayed in Figure 4. The mean and plug-in deviances are shown on the same plot for the AR models, but separate plots are used for the RE models, where the difference between the two deviances is much larger, to maintain consistent scales.

Figure 4: Deviance plots for checking the adequacy of the Qsample length for the HAMD example



## References

- Brooks, S. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–55.
- Celeux, G., Forbes, F., Robert, C. P., and Titterton, D. M. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, **1**, (4), 651–74.
- Daniels, M. J. and Hogan, J. W. (2008). *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall.
- Dempster, A. P. (1973). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference*, (ed. O. Barndorff-Nielsen, P. Blaesild, and G. Schou), pp. 335–54. University of Aarhus.
- Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, **7**, 247–52.
- Diggle, P. and Kenward, M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **43**, (1), 49–93.
- Fitzmaurice, G. M. (2003). Methods for Handling Dropouts in Longitudinal Clinical Trials. *Statistica Neerlandica*, **57**, (1), 75–99.
- Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, (2), 1–19.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*, (2nd edn). Chapman & Hall.
- Jackson, C., Best, N., and Richardson, S. (2006). Improving ecological inference using individual-level data. *Statistics in Medicine*, **25**, 2136–59.
- Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, **8**, (1), 51–83.
- Little, R. J. A. and Rubin, D. B. (1983). On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood. *The American Statistician*, **37**, (3), 218–20.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, (1st edn). Springer-Verlag.
- Michiels, B., Molenberghs, G., Bijmens, L., Vangeneugden, T., and Thijs, H. (2002). Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. *Statistics in Medicine*, **21**, 1023–41.
- Peruggia, M. (1997). On the Variability of Case-Deletion Importance Sampling Weights in the Bayesian Linear Model. *Journal of the American Statistical Association*, **437**, (92), 199–207.
- Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, (2), 147–77.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **64**, (4), 583–639.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge, Available from [www.mrc-bsu.cam.ac.uk/bugs](http://www.mrc-bsu.cam.ac.uk/bugs).
- Wakefield, J. (2004). Ecological inference for 2 x 2 tables. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **167**, (3), 385–445.
- Yun, S.-C., Lee, Y., and Kenward, M. G. (2007). Using hierarchical likelihood for missing data problems. *Biometrika*, **94**, (4), 905–19.