

Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses

Alexina Mason, Sylvia Richardson and Nicky Best

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. This strategy is intended for use within a sensitivity analysis that explores the impact of different assumptions about the two parts of the model, and is illustrated by examples 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.

Keywords: Bayesian model comparison, deviance, DIC, missing response, non-ignorable missingness, observed data likelihood, selection models, sensitivity analysis

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) (henceforth SBCV) 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) (henceforth CRFT) 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 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 with non-ignorable missingness respectively, emphasising that this should be carried out within the context of a sensitivity analysis rather than to select a single ‘best’ model. We conclude 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

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

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

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

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

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

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

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 (SBCV), 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 SBCV, and the posterior mode, which was considered as an alternative by CRFT.

Further, in complex 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 SBCV refer to as the *focus* of the model, and is operationalised by the prediction problem of interest. For example, in a random effects model, if interest lies in models that give good predictions for the observed units or clusters, then the random effects should be included in the model focus. If interest lies in the population parameters and models that give good predictions for new or 'typical' units, then the random effects should not be included in the model focus but integrated out of the likelihood (see also the discussion of model focus in Vaida and Blanchard (2005)).

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 $\bar{D}(\hat{\phi})$ to be the posterior mean of $-2\log L(\phi|\mathbf{y})$, and evaluating $D(\hat{\phi})$ as -2 times the log likelihood at the posterior mean of the stochastic terms in the likelihood. However, other values of DIC can be obtained by using different plug-ins or a different model focus.

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 DIC based on the observed data likelihood

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

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

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:

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

So the first term in the likelihood is the \mathbf{y}_{obs} part of the model of interest, $f(\mathbf{y}_{obs}|\boldsymbol{\lambda})$, and the second term evaluates the model of missingness by integrating over \mathbf{y}_{mis} . The calculation of the expectation in Equation 4 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 fit of the model of interest part to \mathbf{y}_{obs} necessarily deteriorates. This is because, in a selection model, the same model of interest is applied to both the observed and the missing data, and so the $\boldsymbol{\lambda}$ estimates will depend on both \mathbf{y}_{obs} and the imputed \mathbf{y}_{mis} . Since the latter are systematically different from the former under an informative missingness model, the fit of the model of interest *to the observed data* will necessarily be worse than if $\boldsymbol{\lambda}$ had just been estimated using \mathbf{y}_{obs} . Consequently, the value of the part of the DIC_O attributable to the fit of the model of interest will increase when the model of missingness departs from missing at random (MAR). This may partially or completely offset any reduction in the part of DIC_O attributable to improvements in the fit of the model of missingness (as happens in our simulated examples in Section 4 and application in Section 5). Thus, while DIC_O will indicate which selection model best fits the observed data (\mathbf{y}_{obs} and \mathbf{m}), it can be misleading when our true purpose is to compare the fit of selection models to both the observed and missing data (\mathbf{y}_{obs} , \mathbf{m} and \mathbf{y}_{mis}). Neither DIC_O nor any other model selection criterion can answer this question directly as they can never provide information about the fit to the missing data. However, DIC_O can provide useful insight into the comparative fit of certain aspects of these types of models. As we will show, reasonable comparisons can be made using DIC_O by fixing the model of missingness part and using it to compare selection models with different models of interest (conditional on the appropriateness of the missingness model). 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.

Because of the fact that the imputed \mathbf{y}_{mis} will depend on the model of missingness, and DIC_O does not account for the fit to the missing data, we do not recommend using DIC_O to compare selection models with different models of missingness. Hence, it would be useful to have an additional model comparison measure that focusses on the missing data. Clearly we cannot examine the fit of the model to the missing data as we can for the observed data, but we do have information brought by the missingness indicator. We would therefore like a DIC construction that allows us to use this additional information and consider the fit of the model of missingness separately.

2.4 Conditional DIC

An alternative option is a conditional DIC, which treats the missing data as additional parameters (CRFT). This can be written as:

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

For selection models the likelihood on which this is based is

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

For many examples, including all those discussed in this paper, $f(\mathbf{y}_{obs} | \mathbf{y}_{mis}, \boldsymbol{\lambda})$ can be simplified to $f(\mathbf{y}_{obs} | \boldsymbol{\lambda})$. In this case DIC_C only differs from DIC_O in the second term, $f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$, which is evaluated by conditioning on \mathbf{y}_{mis} rather than by integrating over it. The plug-ins for DIC_C include the missing data, \mathbf{y}_{mis} , and can be evaluated as $E(\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})$.

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, and sums these to provide the overall DIC_C . This DIC_C takes the missing observations as part of the model focus, rather than integrating them out as in DIC_O . As a consequence, this overall DIC_C does not focus on the appropriate prediction problem, since, in general, the missing data (and the units for which they arise) are not of direct interest in the joint selection model. Hence we do not recommend using the overall DIC_C calculated by WinBUGS for evaluating selection models. However, if we just consider the model of missingness, then the focus *is* on predicting the missingness for the sampled units in the dataset. The model of missingness part of DIC_C treats \mathbf{y}_{mis} as extra parameters in the model of missingness, with the model of interest acting as their prior distribution, which seems a natural construction for considering the fit of the model of missingness separately. Thus, we propose that the part of DIC_C relating to $f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$, can be used for comparing the fit of different models of missingness for selection models with the same model of interest. While there is no sense in considering this measure in isolation, we will see that it can provide useful additional information when used in conjunction with DIC_O in the context of a sensitivity analysis.

CRFT suggest a different formulation for a conditional DIC, which they call DIC_8 , whereby the missing data are dealt with as missing variables rather than as additional parameters. The idea of DIC_8 is to first condition on the imputed \mathbf{y}_{mis} , calculating the parameters of the model of interest and the model of missingness for each completed dataset, denoted $\hat{\boldsymbol{\lambda}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis})$. Then integrate over \mathbf{y}_{mis} conditional on the observed data $(\mathbf{y}_{obs}, \mathbf{m})$ by averaging the resulting log likelihoods for these datasets. It can be written as:

$$\begin{aligned} \text{DIC}_8 &= 2E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}}\{\log L(\hat{\boldsymbol{\lambda}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis}), \hat{\boldsymbol{\theta}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis}), \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})\} \\ &\quad - 4E_{\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m}}\{\log L(\boldsymbol{\lambda}, \boldsymbol{\theta}, \mathbf{y}_{mis} | \mathbf{y}_{obs}, \mathbf{m})\}. \end{aligned}$$

DIC_8 differs from DIC_C in the plug-in part of the formula (first term), and cannot be computed using only WinBUGS.

Although DIC_8 is a conditional DIC, it does not have a clearly defined focus like DIC_C (or DIC_O), but instead first includes the missing data in the model focus, and then integrates over \mathbf{y}_{mis} . We argue that the resulting ambiguity of the prediction problem that DIC_8 addresses makes it difficult to recommend as a criterion for evaluating the fit of selection models. However, at the request of a referee, we do consider a ‘partial’ DIC_8 just for the model of missingness part of the model likelihood ($f(\mathbf{m}|\mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta})$) and compare this with our recommended partial DIC_C for evaluating the fit of different models of missingness for selection models with the same model of interest.

2.5 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.

3 Implementation of DIC_O and DIC_C

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 this 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. We also discuss some technical issues concerning the choice of plug-ins for calculating both DIC_O and DIC_C .

3.1 Algorithm

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 is 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 the checks that we consider necessary to ensure that the samples

generated for its calculation are of sufficient length.

Our preferred algorithm, used to calculate the DIC_O for selection models implemented in the 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 4, $E_{\mathbf{y}_{mis} | \mathbf{y}_{obs}, \boldsymbol{\beta}^{(k)}, \sigma^{(k)}} \{f(\mathbf{m} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \boldsymbol{\theta}^{(k)})\}$, 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 are 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

$$ESS = \frac{Q}{1 + var(w)} \quad (6)$$

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

$$\frac{\sum_{q=1}^Q (w_q - \bar{w})^2}{(Q - 1)\bar{w}^2} \quad (7)$$

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 (SBCV). (These results are shown in Table 11 of the supplementary material, Section 1.1.) 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 3.2.

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 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 Qsample 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 Qsample. (DIC_O could also be plotted against Qsample 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, presented in Section 4.1 below, 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 Qsamples of length 20,000, 4 non-overlapping Qsamples of length 10,000, 8 non-overlapping Qsamples of length 5,000 and 16 non-overlapping Qsamples 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 that decreases as Q increases.

A similar downwards trend to that seen in \bar{D} , converging to a limit, is indicated by the mean values of \hat{D} .

We consider the Q 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 DIC_O for the model shown in the left panel of Figure 1, but a higher Q might produce a more accurate DIC_O for the models depicted in the middle and right panels of Figure 1. The plots for this and other synthetic examples suggest that higher variability and slower convergence to a limit are associated with poorer fitting models.

3.3 Plug-ins for calculating DIC

Calculation of any DIC involves calculation of a plug-in likelihood \hat{D} evaluated at a point estimate of the model parameters that are in focus, and it is implicitly assumed that this point estimate is a ‘good’ estimate. In situations where the posterior distribution of the parameters in focus is approximately normal and reasonably precise, DIC is likely to be robust to different choices of plug-ins. However, considerable care is needed when calculating the plug-in deviance in situations where the parameters in focus are poorly identified (for example, due to small sample size and weak or conflicting prior information), since this can lead to difficulties in choosing a good point estimate. This is a potential problem for our DIC_C , since this includes the missing data as parameters in the model focus, and these may be only weakly identifiable. We therefore recommend using a number of alternative plug-ins to calculate DIC_C for the missingness model, and examining stability of the resulting DIC_C values to different choices. Plug-ins that result in negative values for p_D , or lead to DIC_C values that are inconsistent with those based on alternative plug-ins, are indicative of problems, and the DIC_C for such models should be interpreted with extreme caution.

In particular, we consider four different plug-ins for calculating DIC_C for the model of missingness in our examples. Let the model of missingness be defined as

$$\begin{aligned} m_i &\sim \text{Bernoulli}(p_i), \\ \text{logit}(p_i) &= f(y_i, \boldsymbol{\theta}), \\ \boldsymbol{\theta} &\sim \text{prior distribution.} \end{aligned} \tag{8}$$

We use (1) the posterior means or (2) the posterior medians of $(\boldsymbol{\theta}, \mathbf{y}_{mis})$ as our plug-in values (which we term “standard” plug-ins) or alternatively (3) the posterior mean or (4) the posterior median of $\text{logit}(p_i)$ as plug-ins (which we term “link” plug-ins). We carry out a number of checks concerning the appropriateness of these plug-ins for our examples, which can be found in the supplementary material, Section 1.3. Results reported for DIC_C in the main paper are based on the two alternative (i.e. *standard* or *link*) posterior median plug-ins; for comparison, results based on the posterior mean plug-ins for DIC_C are provided in the supplementary material, Section 1.6.

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 *standard* plug-ins allow us to evaluate \hat{D} by integrating over \mathbf{y}_{mis} in the model of missingness part of the joint likelihood as required. By contrast, the *link* 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 for the model of missingness part of DIC_O as the one used for calculating DIC_C . Hence for DIC_O , we only use posterior mean *standard* plug-ins, evaluated as $E(\lambda_k|\mathbf{y}_{obs}, \mathbf{m})$ and $E(\theta_k|\mathbf{y}_{obs}, \mathbf{m})$. (Note that posterior median *standard* plug-ins gave virtually identical values of DIC_O and so are not reported here).

4 Illustration of strategy on simulated data

In this section we illustrate our proposed strategy using simulated bivariate Normal data, and demonstrate the limitations of using only DIC_O for comparing selection models with simulated time series data.

4.1 Bivariate Normal data

We now assess how DIC_O and the missingness part of DIC_C can be used to help compare 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.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right). \quad (9)$$

For this dataset the correct model of interest is

$$\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned} \quad (10)$$

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 $\text{logit}(p_i) = -2 + y_i$, which imposes non-ignorable missingness such that the probability of being missing increases with the value of y , and ensures that the estimated probabilities always lie in the range $[0,1]$. This results in 30% of the responses being missing.

Our investigation is based on fitting six joint models (JM1-JM6), as specified in Table 1, to this simulated dataset with simulated missingness. For JM1, both parts of the model are correctly specified. However, JM2 has an inadequate model of interest, JM3 has an incorrect error distribution, JM4-JM5 have too complex a model of missingness, and JM6 is a version of JM5 with a parameter constrained to the wrong sign. So we consider three different models of interest and four different models of missingness. A full implementation of our proposed strategy would involve fitting a set of joint models

Table 1: Specification of joint models for the bivariate Normal simulated data

Model Name	Model of Interest	Model of Missingness
JM1	$y_i \sim N(\mu_i, \sigma^2); \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM2	$y_i \sim N(\mu_i, \sigma^2); \mu_i = \beta_0$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM3	$y_i \sim t_4(\mu_i, \sigma^2); \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i$
JM4	$y_i \sim N(\mu_i, \sigma^2); \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \theta_0 + \theta_1 y_i + \theta_2 y_i^2$
JM5 ^a	$y_i \sim N(\mu_i, \sigma^2); \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \begin{cases} \theta_0 + \theta_1 y_i & : y_i \leq \gamma \\ \theta_0 + \theta_1 \gamma + \theta_2 (y_i - \gamma) & : y_i > \gamma \end{cases}$
JM6 ^{ab}	$y_i \sim N(\mu_i, \sigma^2); \mu_i = \beta_0 + \beta_1 x_i$	$\text{logit}(p_i) = \begin{cases} \theta_0 + \theta_1 y_i & : y_i \leq \gamma \\ \theta_0 + \theta_1 \gamma + \theta_2 (y_i - \gamma) & : y_i > \gamma \end{cases}$

^a The change point in this piecewise regression, γ , is fixed to 0

^b θ_1 is constrained to be positive and θ_2 is constrained to be negative

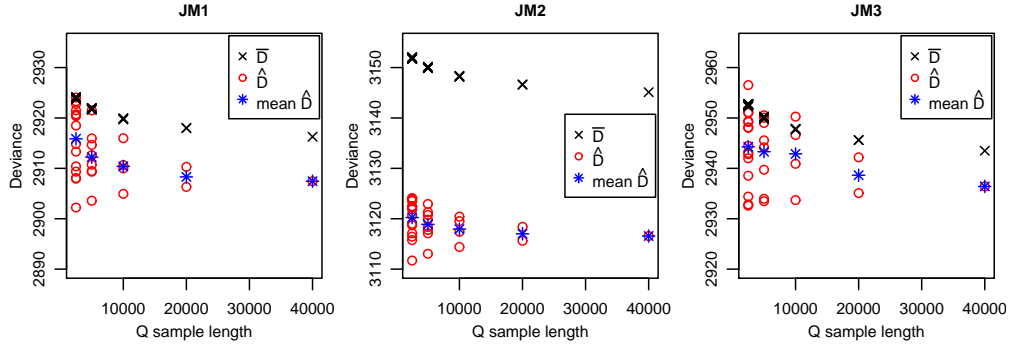
which pairs each model of interest with each model of missingness (twelve joint models), and we do this in our real data application in Section 5.

Vague priors are specified for the unknown parameters of the model of interest: the β 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 θ_0 and a weakly informative $N(0, 0.68)$ prior for θ_1 and θ_2 , which corresponds to an approximately flat prior on the scale of p_i . For JM6, θ_1 is constrained to be positive and θ_2 is constrained to be negative.

We calculate the 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 likelihoods used in the calculations are given in the supplementary material, Section 1.2. The samples produced by the WinBUGS runs are from 2 chains of 25,000 iterations, with 20,000 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 DIC_O for the three models all converged.

As discussed in Section 3.2, Figure 1 allows us to assess the adequacy of the length of our Q 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. In this example, as the DIC_O differ substantially between models, we do not consider it necessary to further reduce the sampling variability by increasing Q any more.

Figure 1: Deviance plots for checking the adequacy of the Qsample 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 DIC_O for JM1-JM3, and two versions of an alternative measure of overall fit, the mean square error (MSE) for the model of interest, as defined by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - E(y_i|\boldsymbol{\beta}))^2, \quad (11)$$

where $E(y_i|\boldsymbol{\beta})$ is evaluated as the posterior mean of μ_i in Equation 10. One version (obs) is based only on the observed data, and the other (all) uses both the observed and missing data which is possible with this dataset because we have simulated the missingness and know the true values of the missing y_i . Based only on the observed data, JM2 is clearly a very poor choice, but there is nothing to choose between JM1 and JM3. If the missing data are also brought into consideration, the message is the same. In contrast to the MSE, DIC_O assesses the joint fit of both parts of the model, penalised for complexity, although, as with MSE obs, the fit of the model of interest is only considered with respect to the observed responses. Following our proposed strategy, we only use DIC_O to compare models with the same model of missingness, i.e. JM1-JM3, and DIC_O correctly suggests that JM1 is a better fitting model than JM2 or JM3. However, if we calculate DIC_O for JM4-JM6 (see Table 2) and use DIC_O to compare models with a different model of missingness, we would conclude that the other three models fit slightly better than JM1.

We now look at the model of missingness part of DIC_C , to compare JM1 and JM4-JM6, the four models with the same model of interest. In calculating 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 using the *standard* plug-ins and the other calculated using *link* plug-ins, and look for broad consistency across the plug-ins. Inconsistency or negative p_D are taken as indicative of model problems. The DIC_C based on *standard* plug-ins returns a negative p_D for JM4 and there is inconsistency between the two versions of DIC_C , so we have reservations about this model. A comparison of the model of missingness DIC_C for the remaining three models (Table 3) suggests clearly that the

Table 2: Comparison of DIC_O and MSE for JM1-JM6

	<i>standard</i> plug-ins				MSE	
	\bar{D}	\hat{D}	p_D	DIC_O	all ^a	obs ^b
JM1	2916.3	2907.4	8.8	2925.1	0.741	0.705
JM2	3145.1	3116.6	28.5	3173.7	0.998	0.884
JM3	2943.5	2936.4	7.1	2950.6	0.740	0.704
JM4	2908.6	2898.8	9.8	2918.4	0.746	0.710
JM5	2913.4	2905.2	8.1	2921.5	0.742	0.706
JM6	2917.2	2912.4	4.9	2922.1	0.772	0.668

^a MSE based on all data (observed and missing)

^b MSE based on observed data only

model of missingness in JM6 provides a much poorer fit than those for JM1 and JM5. There is rather less to choose between JM1 and JM5, suggesting that while DIC_C is useful in identifying substantially wrong models, it is a little too conservative in its penalisation for complexity for distinguishing between a correct model and a slightly over complex version of that model.

Table 3: Model of missingness DIC_C (calculated using posterior median plug-ins) and DIC_8 for the simulated bivariate normal data for JM1 and JM4-JM6

	<i>standard</i> plug-ins				<i>link</i> plug-ins			\hat{D}	p_D	DIC_8
	\bar{D}	\hat{D}	p_D	DIC_C	\hat{D}	p_D	DIC_C			
JM1	1072.3	1032.2	40.2	1112.5	1033.3	39.1	1111.4	1070.4	2.0	1074.3
JM4	1060.3	1065.8	-5.5	1054.8	1050.6	9.7	1070.0	1057.3	3.0	1063.3
JM5	1069.3	1032.8	36.5	1105.8	1029.5	39.9	1109.2	1066.4	2.9	1072.3
JM6	1212.2	1186.8	25.5	1237.7	1193.5	18.7	1231.0	1208.3	4.0	1216.2

Table 3 also shows the model of missingness DIC_8 , the alternative formulation of a conditional DIC suggested by CRFT, which was introduced in Section 2.4. The posterior mean deviance is as for DIC_C and hence generated by WinBUGS, but some additional calculations are required to produce the plug-in deviance. Firstly, the \mathbf{y}_{mis} sample produced by the WinBUGS MCMC run is used to generate K complete datasets. Then using the R software, the model of missingness logistic regression is fitted to each complete dataset in turn to produce $\hat{\boldsymbol{\theta}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis})$. (We are able to fit the model of missingness separately as the likelihood factorises in the presence of the full data.) Finally, the values of the partial DIC are averaged over the K simulated datasets. Fuller

details of this algorithm are given in the Appendix.

The p_D for DIC_8 is close to the actual number of parameters (not including the missing data) in the model of missingness for this example, which is consistent with CFRT's findings for mixture models. This suggests that, despite conditioning on the missing data in the first part of its calculation, DIC_8 behaves like a 'population focused' DIC that does not penalize at all the imputation of the missing data. The ordering of the partial DIC_8 resembles DIC_C , and suggests that we should favour JM4, as there are no clear warnings from a negative p_D or inconsistency. The appropriateness of this alternative message depends on the question that we are trying to answer. However, in general, the DIC_C , which focuses on identifying models that provide good predictions of the missingness mechanism for the current set of sampling units, seems more appropriate for our purposes than DIC_8 , which focuses on prediction for other sampling units subject to the same missingness mechanism.

Summarising, if we use DIC_O to compare models with the same model of missingness, we will prefer JM1 to JM2 and JM3, and if we use the partial DIC_C to compare different missingness models with the same model of interest we will find JM1 and JM5 to be the most plausible models. These two models give identical estimates of β_0 , 1.02 (0.94,1.11), and very similar estimates of β_1 : 0.53 (0.46,0.60) from JM1 and 0.55 (0.48,0.63) from JM5 (posterior mean with 95% interval in brackets). As there is little to choose between JM1 and JM5 in terms of performance and DIC_C is inconclusive, we follow the principle of parsimony and favour JM1. So in this example, a combination of DIC_O , DIC_C and performance (correctly) point towards JM1 being the best model.

We now consider some simulated longitudinal data which mimics the basic structure of the clinical trial data that we will analyse in Section 5, in order to illustrate why DIC_O can be misleading if used in isolation, and should not be used to compare selection models with different models of missingness.

4.2 Time series data

For our second simulation we generate response data, y_{it} , for $i = 1, \dots, 1000$ individuals at two time points, $t = 1, 2$, using the random effects model:

$$\begin{aligned} y_{it} &\sim N(\mu_{it}, \sigma^2) \\ \mu_{it} &= \beta_i + \eta t \\ \beta_i &\sim N(\gamma, \rho^2) \end{aligned} \tag{12}$$

with $\sigma = 1$, $\eta = -1$, $\gamma = 0$ and $\rho = 1$. We then impose non-ignorable missingness on y_{i2} according to the linear logistic equation, $\text{logit}(p_i) = y_{i2} - y_{i1}$, where p_i is the probability that y_{i2} is missing. So in this example, the missingness is dependent on the change in y_i between time points. Three joint models are fitted to this data, all with a correctly specified model of interest, as given by Equation 12, but different models of missingness as specified in Table 4. The priors are similar to those specified for the models in Section 4.1, and all the models exhibit satisfactory convergence. The model of interest parameter estimates show that mmar2 is closest to fitting the true

data generating model (see Table 15 in the supplementary material, Section 1.4).

Table 4: Specification of the models of missingness for the simulated time series data

Model Name	Model of Missingness equation
mar	$\text{logit}(p_{iw}) = \theta_0 + \theta_1 y_{i1}$
mnar	$\text{logit}(p_{iw}) = \theta_0 + \theta_2 y_{i2}$
mnar2	$\text{logit}(p_{iw}) = \theta_0 + \theta_3 y_{i1} + \theta_4 (y_{i2} - y_{i1})$

We calculate DIC_O for the three models as described in the previous section. A Q_{sample} length of 40,000 is adequate (see Figure 4 in the supplementary material, Section 1.7). In Section 2.3 we discussed the limitations of DIC_O for comparing selection models with different models of missingness and recommended that DIC_O is not used for this purpose. Table 5 provides empirical support for this recommendation, as the correct model (mnar2) clearly has the highest DIC_O . Although the higher value of \hat{D} for mnar2 compared with mar and mnar may seem counter intuitive, it reflects the discussion in Section 2.3 about the deterioration in fit of the model of interest to the observed data when a non-ignorable non-response mechanism is assumed (see supplementary material, Section 1.5 for further detail). However, if instead we use the model of missingness part of DIC_C , in line with our proposed strategy, we will conclude that the mnar2 model best explains the missingness pattern regardless of the plug-ins chosen (Table 6). The model of missingness part of DIC_8 also suggests that we should favour mnar2.

Table 5: DIC_O for the simulated time series data

	\bar{D}	<i>standard</i> plug-ins		
		\hat{D}	p_D	DIC_O
mar	5515.4	4788.3	727.1	6242.5
mnar	5462.7	4721.2	741.5	6204.1
mnar2	6004.0	5382.7	621.2	6625.2

These findings were replicated with two further datasets, randomly generated using the same equations, and support our proposed two-measure strategy in preference to using a single DIC measure for comparing selection models with suspected non-ignorable missing responses.

We now examine this approach in a case study comparing three treatments of depression using longitudinal data.

Table 6: Model of missingness DIC_C and DIC_8 for the simulated time series data

	<i>standard</i> plug-ins				<i>link</i> plug-ins				DIC_8	
	\bar{D}	\hat{D}	p_D	DIC_C	\hat{D}	p_D	DIC_C	\hat{D}		p_D
mar	1198.0	1196.1	1.9	1199.9	1196.1	2.0	1200.0	1196.1	1.9	1199.9
mnar	1089.9	1051.5	38.4	1128.3	1051.5	38.4	1128.2	1087.8	2.1	1092.0
mnar2	813.7	675.4	138.4	952.1	677.6	136.2	949.9	810.7	3.0	816.7

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). DK found evidence of informative missingness given their modelling assumptions. 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.

The missing responses in the HAMD data force us to make modelling assumptions that are untestable from the data, and no measure can tell the whole story regarding model fit. In these circumstances we know that sensitivity analysis is essential, and cover a range of options by proposing two different models of interest and three different models of missingness. Attempting to select a single ‘best’ model from the six possible combinations would defeat the object of the sensitivity analysis, but we use our two measure DIC strategy to help determine whether some of these models are more plausible than others.

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 δ_{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.

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

The parameters capturing the treatment effects are $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, and the treatment effects will be displayed graphically. 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 $\text{Gamma}(0.001,0.001)$ 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 7, and assign a logistic prior to θ_0 and weakly informative Normal priors to all the other $\boldsymbol{\theta}$ parameters as previously discussed (Section 4.1). 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 θ_3 . MoM3 has the same form as MoM2, but includes separate $\boldsymbol{\theta}$ for each treatment, which allows treatment to directly affect the missingness process.

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

Table 7: 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	$\text{logit}(p_{iw}) = \theta_{0t(i)} + \theta_{2t(i)} y_{i(w-1)} + \theta_{3t(i)} (y_{iw} - y_{i(w-1)})$

based on the Gelman-Rubin convergence statistic and a visual inspection of the chains. Adding a missingness model makes little difference to the β or γ estimates, but there are substantial changes in some of the η and ξ 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 8. The positive θ_1 estimates for the JM1 models suggest that drop-out is associated with high HAMD scores, while the negative θ_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 8: Parameter estimates for the model of missingness for the HAMD data

	JM1(AR)	JM2(AR)	JM3(AR)	JM1(RE)	JM2(RE)	JM3(RE)
θ_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)
θ_1	0.08 (0.04,0.11)			0.04 (0.01,0.08)		
θ_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)
θ_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)

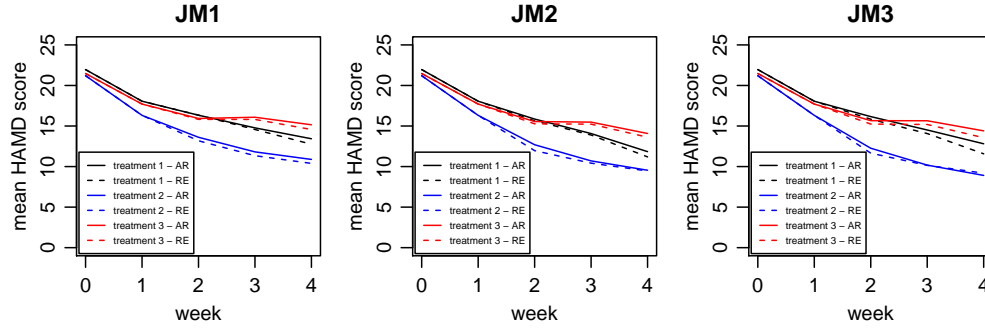
Table shows the posterior mean, with the 95% interval in brackets

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

DH point out that correctly specifying the dependence structure in the model of interest has increased importance when dealing with missing data. 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 JM1 and JM2 the solid (AR) and dashed (RE) lines for each treatment show very small differences, which accentuate slightly for the more complex JM3.

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



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

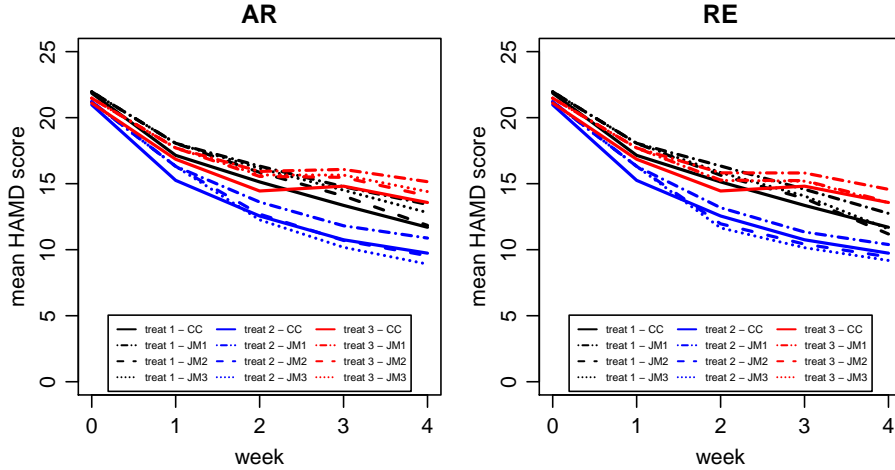
The impact of the addition of the MoM1 model of missingness to the AR model of interest can be seen by comparing the CC (solid lines) and JM1 (dot-dash lines) in Figure 3 and noticing a small upward shift of JM1; the impact is slightly less when the RE model of interest is used. By contrast, the direction and magnitude of the shift from CC varies according to treatment and week when either MoM2 (dashed lines) or MoM3 (dotted lines) is added to either model of interest.

5.5 Use of DIC_O to help with model comparison

DIC_O is calculated for the six HAMD models using the algorithm discussed in Section 3. 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. The Ksample length 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 9 shows DIC_O for the six models 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.

Before discussing these results, we examine the adequacy of the Q_{sample} , by splitting it into subsets and plotting \bar{D} and \hat{D} against the sample lengths as described in Section 3. From these plots for the six models (shown as Figure 5 in the supplementary material, Section 1.8), 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

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



In the RE plot, the JM2 and JM3 lines for treatment 3 are almost coincident.

Table 9: DIC_O for the HAMD data

	\bar{D}	<i>standard plug-ins</i>		
		\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

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 the relative merits of the model of interest. 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 in turn, 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 comparison

We now turn to the model of missingness part of DIC_C , to see what additional information it provides. The two versions shown in Table 10, based on the *standard* plug-ins and the *link* plug-ins, provide a consistent message. For a given model of interest (AR or RE) MoM2 and MoM3, used in the JM2 and JM3 models, clearly provide a better fit to this part of the model than JM1, with evidence that JM3 is preferable to JM2, i.e. a missingness model that allows treatment specific parameters.

Table 10: Model of Missingness DIC_C (calculated using posterior median plug-ins) for the HAMD data

	\bar{D}	<i>standard</i> plug-ins			<i>link</i> plug-ins		
		\hat{D}	p_D	DIC_C	\hat{D}	p_D	DIC_C
JM1(AR)	698.6	695.5	3.1	701.7	695.9	2.7	701.3
JM2(AR)	653.4	648.4	5.0	658.3	652.5	0.9	654.3
JM3(AR)	626.0	617.8	8.2	634.2	607.4	18.6	644.6
JM1(RE)	719.6	717.6	2.0	721.7	718.4	1.3	720.9
JM2(RE)	547.5	514.6	32.9	580.4	516.3	31.2	578.7
JM3(RE)	521.6	478.0	43.6	565.2	480.9	40.7	562.4

5.7 Combined use of DIC_O and the model of missingness DIC_C

To conclude, within this sensitivity analysis, DIC_O suggests that the RE model of interest is more plausible 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. If we wished to report the results of a single model, then JM3(RE) would be the best option. However, JM2(RE) is also reasonably well supported, so in the spirit of sensitivity analysis, we should report results from both models with a RE model of interest and a model of missingness that depends on the change in HAMD (either treatment specific or not). The results from these two models are robust.

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. The same conclusions are reached using our preferred joint

models, i.e. JM2(RE) and JM3(RE), but by the end of the study, treatment 2 appears a little more effective in lowering HAMD (compare the dotted 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. No single measure of DIC, or indeed combination of measures, can provide a full picture of model fit since we can never evaluate fit to the missing data. However, the use of two complementary measures can provide more information than one DIC measure used in isolation. The model comparison strategy that we have developed relies on using both DIC_O and the model of missingness part of DIC_C . 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. The model of missingness part of DIC_C , which uses information provided by the missingness indicators, allows comparison of the fit of different models of missingness for selection models with the same model of interest. In view of the difficulty of choosing plug-ins that provide robust estimates of DIC_C , we recommend that different plug-ins are used and inconsistency interpreted as flagging an unreliable model.

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_{samples}) 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_{sample} of at least 40,000. Again based on our experience, we tentatively suggest that even with a well chosen Q_{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.

A model's fit to the observed data can be assessed, but its fit to the unobserved data given the observed data cannot be assessed. So, 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, $L(\boldsymbol{\beta}, \boldsymbol{\theta} | \mathbf{y}_{obs}, \mathbf{y}_{mis}, \mathbf{m})$, and provided a broad outline for its implementation. DIC_F may provide additional information for model comparison, but its calculation is complicated as the expectation for the plug-ins is conditional on \mathbf{y}_{mis} . We have found it to be computationally very unstable in preliminary investigations (DH also noted similar computational problems; personal communication).

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. In surveys, sometimes data is collected from individuals who are originally non-contacts or refusals, and using this for comparing model fit is particularly attractive as such individuals are likely to be similar to those who have missing data. By contrast, alternatives such as K-fold validation will only tell us about the fit to the observed data and as such provide an alternative to the DIC_O part of the strategy. The link between cross-validation and DIC is discussed by Plummer (2008).

Although the DIC_O and model of missingness DIC_C 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. Rather our strategy should be viewed as a screening method that can help us to identify plausible models. Even with straightforward data, such as our first 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, we emphasise that DIC should never be used in isolation. Our DIC strategy should be used in the context of a sensitivity analysis, designed to check that conclusions are robust to a range of assumptions about the missing data. In summary, our investigations have shown that these two DIC measures have the potential to assist in the selection of a range of plausible models which have a reasonable fit to quantities that can be checked and allow the uncertainty introduced by non-ignorable missing data to be propagated into conclusions about a question of interest.

Appendix

Algorithm for calculating DIC_O

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 $\boldsymbol{\beta}$, σ 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 $\boldsymbol{\beta}$, σ 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 in Section 1.3 of 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 $Q_{sample}^{(k)}$.
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.

Algorithm for calculating the plug-in deviance for the model of missingness part of DIC_8

1. Carry out a standard MCMC run on the joint model $f(\mathbf{y}, \mathbf{m}|\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$. Save samples of \mathbf{y}_{mis} , denoted by $\mathbf{y}_{mis}^{(k)}$, $k = 1, \dots, K$, and use these to form K complete datasets.
2. Fit the model of missingness part of the joint model, $f(\mathbf{m}|\mathbf{y}, \boldsymbol{\theta})$, to each complete dataset to calculate $\hat{\boldsymbol{\theta}}(\mathbf{y}_{obs}, \mathbf{m}, \mathbf{y}_{mis})$.
3. Then average results from the K datasets to get the plug-in log likelihood:

$$\begin{aligned} & E_{\mathbf{y}_{mis}|\mathbf{y}_{obs},\mathbf{m}}\{\log f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis},\hat{\boldsymbol{\theta}}(\mathbf{y}_{obs},\mathbf{m},\mathbf{y}_{mis}))\} \\ & \approx \frac{1}{K} \sum_{k=1}^K \left[\log f(\mathbf{m}|\mathbf{y}_{obs},\mathbf{y}_{mis}^{(k)},\hat{\boldsymbol{\theta}}(\mathbf{y}_{obs},\mathbf{m},\mathbf{y}_{mis}^{(k)})) \right]. \end{aligned}$$

Multiply this plug-in log likelihood by -2 to get the plug-in deviance.

1 Supplementary Material

1.1 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 11.

Table 11: 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 ^a	2418.1	2410.8	7.3	2425.4
Repetition2a - Ksample seed changed ^a	2418.2	2411.0	7.2	2425.4
Repetition3a - Ksample seed changed ^a	2418.1	2410.9	7.2	2425.3
Repetition4a - Ksample seed changed ^a	2418.3	2411.0	7.4	2425.7
Repetition1b - Qsample seed changed ^b	2419.1	2410.4	8.8	2427.9
Repetition2b - Qsample seed changed ^b	2420.0	2413.5	6.4	2426.4
Repetition3b - Qsample seed changed ^b	2423.0	2416.0	7.0	2430.0
Repetition4b - Qsample seed changed ^b	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.

^a Each repetition uses a different random number seed to generate the Ksample, but the same random number seed to generate the Qsample.

^b Each repetition uses the same random number seed to generate the Ksample, but a different random number seed to generate the Qsample.

1.2 The likelihood equations in the DIC_O calculations for the simulated bivariate Normal example

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

calculated using

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

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

1.3 Skewness in the plug-ins

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

$$\text{coefficient of skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\text{sd}(x)^3}, \quad (16)$$

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

Simulated bivariate Normal 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 bivariate Normal example are shown in Table 12. Mean and 95% interval values are given for \mathbf{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. So, some caution is required in interpreting the DIC_O , and the use of posterior medians rather than posterior means as our plug-ins for calculating DIC_C seems prudent.

σ , $\log(\sigma)$ and τ 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 σ is appropriate as argued by SBCV. (All our DIC calculations work with plug-in values for σ calculated on the log scale.)

Table 12: Skewness of posterior distribution of plug-ins for the simulated bivariate Normal data (skewness outside the interval (-1,1) highlighted in bold)

	JM1	JM2	JM3	JM4	JM5	JM6
β_0	0.09	-0.23	0.00	0.19	0.11	-0.08
β_1	0.10		0.08	0.30	0.29	0.05
σ	0.30	0.72	0.20	0.33	0.34	0.09
$\log(\sigma)$	0.19	0.59	0.09	0.21	0.22	0.01
τ	0.01	-0.34	0.14	0.04	-0.01	0.15
θ_0	-0.47	-1.13	-0.45	-0.87	-0.16	0.10
θ_1	0.27	-0.16	0.19	0.98	0.16	0.15
θ_2				-0.29	0.08	-1.94
y_{mis}^a	-0.01	-0.06	1.83	-0.20	-0.12	0.13
	(-0.11,0.08)	(-0.14,0.01)	(0.90,3.83)	(-0.33,-0.07)	(-0.32,0.08)	(0.02,0.26)
$\text{logitp}(\text{all})^a$	-0.19	-0.34	0.40	-0.05	-0.09	-1.19
	(-0.48,0.25)	(-1.58,1.18)	(-0.45,2.49)	(-1.60,1.44)	(-0.43,0.35)	(-4.80,0.08)
$\text{logitp}(\text{obs})^a$	-0.33	-0.94	-0.31	-0.61	-0.24	-0.07
	(-0.48,-0.03)	(-1.61,-0.27)	(-0.45,-0.04)	(-1.63,-0.02)	(-0.46,-0.05)	(-0.30,0.11)
$\text{logitp}(\text{mis})^a$	0.13	1.04	2.03	1.24	0.26	-3.74
	(-0.08,0.32)	(0.82,1.28)	(1.06,3.89)	(0.96,1.63)	(0.12,0.42)	(-5.64,-1.32)

^a mean value is shown, with 95% interval below in brackets

Simulated time series 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 time series data are shown in Table 13. The skewness for this example is generally less problematic than for the bivariate Normal example.

Table 13: Skewness of posterior distribution of plug-ins for the simulated time series data

	mar	mnar	mnar2
η	0.097	0.105	-0.207
σ	0.143	0.126	-0.029
$\log(\sigma)$	0.060	0.047	-0.135
τ	0.103	0.110	0.350
γ	-0.040	-0.138	0.129
ρ	0.075	0.068	-0.045
θ_0	-0.103	-0.578	0.049
θ_1	-0.091		
θ_2		-0.499	
θ_3			0.249
θ_4			0.283
y_{mis}^a	-0.002 (-0.090,0.083)	0.009 (-0.114,0.122)	0.050 (-0.121,0.205)
$\text{logitp}(\text{all})^a$	-0.089 (-0.124,-0.031)	-0.199 (-0.587,0.235)	-0.151 (-0.450,0.352)
$\text{logitp}(\text{obs})^a$	-0.093 (-0.125,-0.040)	-0.340 (-0.587,0.189)	-0.336 (-0.458,0.071)
$\text{logitp}(\text{mis})^a$	-0.081 (-0.122,-0.017)	0.090 (-0.100,0.287)	0.230 (0.047,0.417)

^a mean value is shown, with 95% interval below in brackets

HAMD example

In our application, we find from looking at the coefficients of skewness for the posterior distributions of the plug-ins (Table 14), that some are skewed, most notably for the two JM3 models.

Table 14: Skewness of posterior distributions of plug-ins for the HAMD example

	JM1(AR)	JM2(AR)	JM3(AR)	JM1(RE)	JM2(RE)	JM3(RE)
β_1/γ_1^a	0.01	-0.03	-0.05	-0.01	-0.06	-0.05
β_2/γ_2^a	-0.04	0.08	0.03	0.04	0.01	-0.01
β_3/γ_3^a	0.09	0.04	-0.06	0.01	0.07	0.07
β_4/γ_4^a	-0.07	0.01	0.00	0.09	-0.08	0.01
β_5/γ_5^a	-0.01	0.05	0.03	0.07	0.13	-0.06
β_6/γ_6^a	0.10	0.05	-0.01	0.00	-0.01	0.06
η_1	-0.10	0.05	0.19	-0.07	0.06	0.08
η_2	0.06	0.00	0.06	0.01	0.09	-0.01
η_3	0.02	-0.01	0.02	0.02	0.04	-0.11
ξ_1	0.04	-0.09	-0.11	0.08	-0.09	0.03
ξ_2	-0.03	0.01	-0.04	-0.02	-0.12	0.00
ξ_3	0.00	-0.02	-0.05	0.04	-0.03	0.09
θ_0	-0.06	-0.08		-0.12	-0.16	
$\theta_0(t_1)$			-0.41			-0.21
$\theta_0(t_2)$			-0.69			-0.35
$\theta_0(t_3)$			-0.22			-0.26
θ_1	-0.01			0.00		
θ_2		-0.20			-0.21	
$\theta_2(t_1)$			-0.04			-0.20
$\theta_2(t_2)$			-0.39			-0.90
$\theta_2(t_3)$			-0.09			-0.51
θ_3		-0.05			-0.24	
$\theta_3(t_1)$			0.40			-0.13
$\theta_3(t_2)$			-0.04			-0.76
$\theta_3(t_3)$			-0.28			-0.43
σ	0.11	0.10	0.10	0.16	0.04	0.17
$\log(\sigma)$	0.05	0.04	0.04	0.11	-0.03	0.09
y_{mis}^b	0.00	0.04	0.10	0.00	0.09	0.09
	(-0.09,0.09)	(-0.06,0.14)	(-0.04,0.29)	(-0.09,0.09)	(-0.04,0.20)	(-0.09,0.26)
logitp^b	-0.06	-0.32	-0.57	-0.10	-0.32	-0.52
	(-0.12,0.20)	(-0.72,0.46)	(-1.33,0.67)	(-0.21,0.30)	(-0.55,0.13)	(-1.09,0.42)

^a β for AR2 models and γ for RE models^b mean value is shown, with 95% interval below in brackets

1.4 Parameter estimates for the simulated time series example

Parameter estimates for the model of interest part of the three models fitted to the simulated time series data are shown in Table 15.

Table 15: Model of interest parameter estimates (posterior means, with 95% credible intervals in brackets) for the simulated time series data

	actual	mar		mnar		mnar2	
σ	1	0.88	(0.83,0.93)	0.89	(0.84,0.94)	1.06	(0.99,1.14)
ρ	1	1.14	(1.07,1.21)	1.18	(1.11,1.25)	1.00	(0.92,1.08)
γ	0	0.50	(0.35,0.65)	0.76	(0.59,0.93)	-0.03	(-0.21,0.16)
η	-1	-1.44	(-1.53,-1.35)	-1.70	(-1.81,-1.58)	-0.92	(-1.07,-0.78)

1.5 Breakdown of \bar{D} and \hat{D} from the DIC_O calculations for the simulated time series data

As discussed in Section 2.3, when we allow for informative missingness, the fit of the model of interest to \mathbf{y}_{obs} deteriorates. This is shown for the simulated time series data in Table 16, which provides the two components of the mean deviance and plug-in deviance separately. There are improvements in the fit of the model of missingness part, but for this example they are insufficient to offset the deterioration in the model of interest part.

Table 16: Model of interest and model of missingness contributions to \bar{D} and \hat{D} from the calculation of DIC_O for the simulated time series data

	\bar{D}		\hat{D}	
	moi ^a	mom ^b	moi ^a	mom ^b
mar	4317.3	1198.0	3592.1	1196.1
mnar	4321.7	1141.0	3593.0	1128.2
mnar2	4902.1	1101.9	4333.8	1048.9

^a moi = contribution from \mathbf{y}_{obs} part of model of interest;

^b mom = contribution from model of missingness

1.6 Model of Missingness DIC_C tables calculated using posterior mean plug-ins

Table 17: Model of missingness DIC_C for the simulated bivariate normal data

	\bar{D}	<i>standard</i> plug-ins			<i>link</i> plug-ins		
		\hat{D}	p_D	DIC_C	\hat{D}	p_D	DIC_C
JM1	1072.3	1032.5	39.9	1112.2	1025.9	46.5	1118.8
JM4	1060.3	1077.2	-16.9	1043.4	979.8	80.5	1140.8
JM5	1069.3	1040.2	29.1	1098.5	1016.7	52.6	1122.0
JM6	1212.2	1187.8	24.4	1236.7	1207.8	4.4	1216.7

Table 18: Model of missingness DIC_C for the simulated time series data

	\bar{D}	<i>standard</i> plug-ins			<i>link</i> plug-ins		
		\hat{D}	p_D	DIC_C	\hat{D}	p_D	DIC_C
mar	1198.0	1196.1	1.9	1200.0	1196.1	1.9	1199.9
mnar	1089.9	1050.3	39.6	1129.5	1046.6	43.3	1133.2
correct	813.7	671.6	142.2	955.9	663.3	150.5	964.2

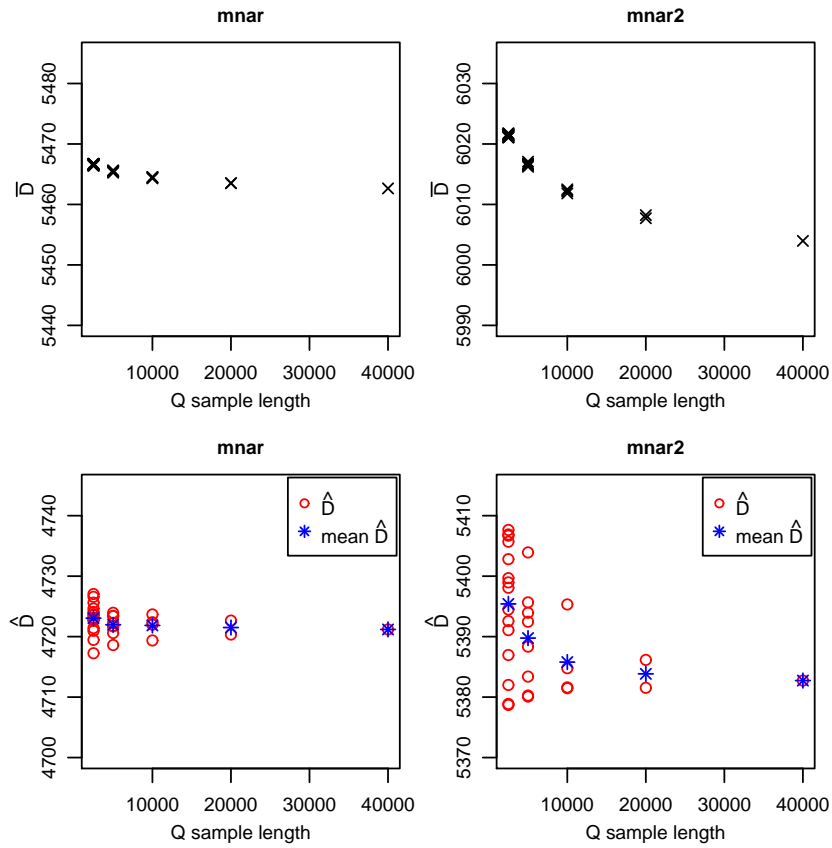
Table 19: Model of missingness DIC_C for the HAMD data

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

1.7 Adequacy of the Qsample length for the simulated times series example

Plots of \bar{D} and \hat{D} against the Qsample lengths for the three models in the time series simulation, as described in Section 4.2, are displayed in Figure 4. Separate plots are used for the mean and plug-in deviances, to more easily assess the convergence towards a limit and stability in \hat{D} . These plots are only shown for the mnar and mnar2 models, since by definition, the model of missingness for the mar model does not contain y_{mis} so there is no need to generate a Qsample.

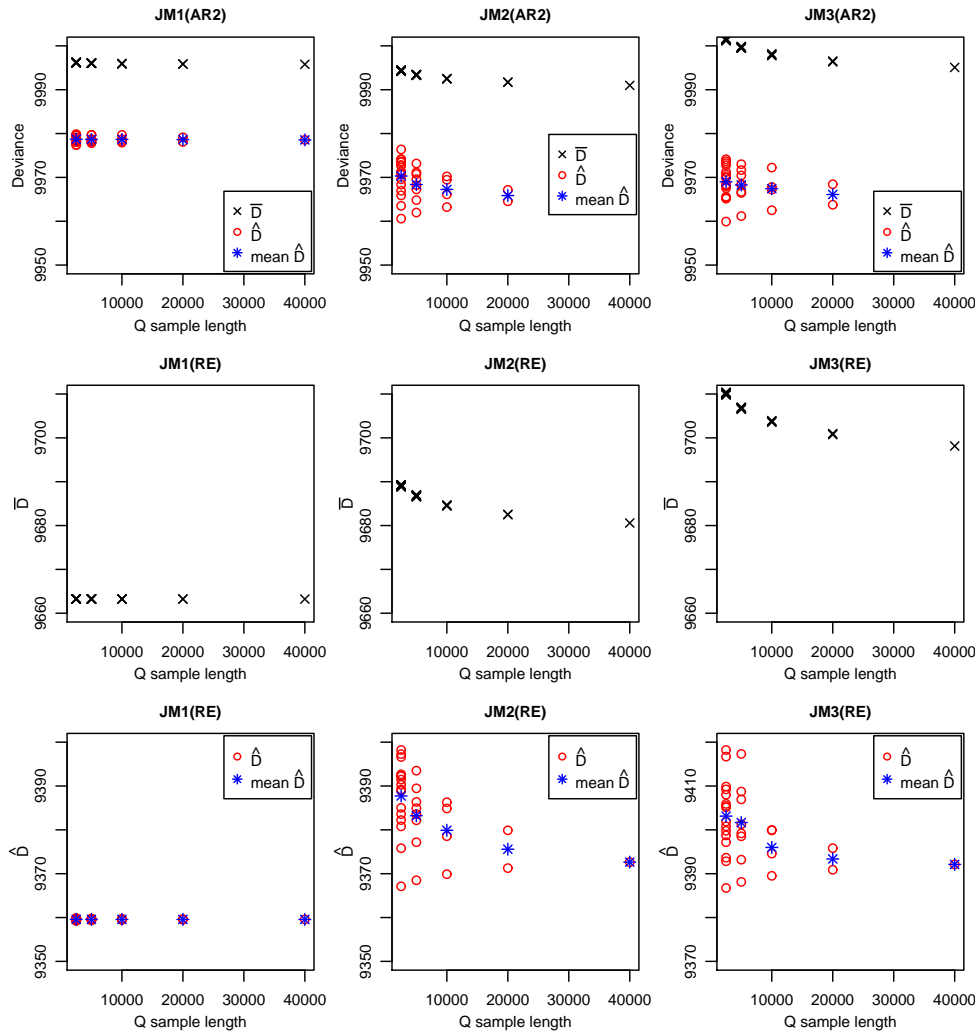
Figure 4: Deviance plots for checking the adequacy of the Qsample length for the simulated time series data



1.8 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 5. 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 5: Deviance plots for checking the adequacy of the Qsample length for the HAMD data



References

- Brooks, S. and Gelman, A. (1998). “General Methods for Monitoring Convergence of Iterative Simulations.” *Journal of Computational and Graphical Statistics*, 7: 434–455.
- Celeux, G., Forbes, F., Robert, C. P., and Titterton, D. M. (2006). “Deviance Information Criteria for Missing Data Models.” *Bayesian Analysis*, 1(4): 651–674.
- 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 Barndorff-Nielsen, O., Blaesild, P., and Schou, G. (eds.), *Proceedings of Conference on Foundational Questions in Statistical Inference*, 335–354. University of Aarhus.
- (1997). “The direct use of likelihood for significance testing.” *Statistics and Computing*, 7: 247–252.
- Diggle, P. and Kenward, M. G. (1994). “Informative Drop-out in Longitudinal Data Analysis (with discussion).” *Journal of the Royal Statistical Society - Series C*, 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*. Chapman & Hall, 2nd edition.
- Jackson, C., Best, N., and Richardson, S. (2006). “Improving ecological inference using individual-level data.” *Statistics in Medicine*, 25: 2136–2159.
- 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.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, 1st edition.
- Michiels, B., Molenberghs, G., Bijns, 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: 1023–1041.
- 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.

- Plummer, M. (2008). "Penalized loss functions for Bayesian model comparison." *Biostatistics*, 9(3): 523–539.
- Schafer, J. L. and Graham, J. W. (2002). "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7(2): 147–177.
- 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*, 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.
- Vaida, F. and Blanchard, S. (2005). "Conditional Akaike information for mixed-effects models." *Biometrika*, 92(2): 351–370.
- Wakefield, J. (2004). "Ecological inference for 2 x 2 tables." *Journal of the Royal Statistical Society - Series A*, 167(3): 385–445.
- Yun, S.-C., Lee, Y., and Kenward, M. G. (2007). "Using hierarchical likelihood for missing data problems." *Biometrika*, 94(4): 905–919.

Acknowledgments

Financial support: this work was supported by an ESRC PhD studentship (Alexina Mason) and ESRC grants: RES-576-25-5003 and RES-576-25-0015. Sylvia Richardson and Nicky Best are investigators in the MRC-HPA centre for environment and health. 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. The authors also thank the Associate Editor and anonymous reviewers for their thoughtful comments.