# Multiple Imputation based Sensitivity Analysis, Ref STAT07852

James R. Carpenter
London School of Hygiene & Tropical Medicine and MRC Clinical Trials Unit at UCL
Email: james.carpenter@lshtm.ac.uk
June 22, 2018

## 1 Overview

In this article, we introduce the concept of sensitivity analyses for missing data, and outline how these may be conveniently performed using Multiple Imputation (MI). To motivate and illustrate the approach, we use data from a trial with missing outcome values, and a cohort study with missing covariate and outcome values. While many applications readers will encounter will be more complex than this, the same concepts will apply and MI — performed following the principles set out here, and the details set out in the references — should provide a practical approach.

The article is structured as follows. We begin by describing the motivating peer review study example in Section 2, which leads to our general framework, of pattern mixture and selection models, in Section 3. We consider pattern mixture modelling using MI in Section 4, first with missing outcomes, and then in Subsection 4.1 with missing covariates. Continuing in the pattern mixture framework, we briefly discuss using expert opinion to inform sensitivity parameters (Subsection 4.2); information anchored approaches (Subsection 4.3) and longitudinal data (Subsection 4.4). We then consider selection models, and discuss an approximate method which is particularly useful when the analyst is not the imputer in Section 5. We conclude with a brief discussion in Section 6.

## 2 Motivating example: peer review study

In response to concerns about the quality of peer review, Schroter *et al.* (2004) carried out a randomised controlled trial to compare face-to-face training of peer reviewers with a self-taught pacakage (using the face-to-face training materials) and no training.

Three previously published BMJ papers were selected, each describing a RCT of alternative, generic, ways of organising and managing clinical work. The names of the original authors were removed, and titles and any reference to study location were changed. Further, 9 major and 5 minor errors were introduced. The primary outcome was the review quality, measured by the *Review Quality Instrument* (RQI) (van Rooyen *et al.*, 1999). Each review was scored independently by two editors, and the outcome was the mean score of the items averaged over the two ratings, which lies between 0 and 5 (excellent).

After consent, but before randomisation, each participating reviewer was sent the first, baseline, paper to review. Reviewers who returned this baseline paper were randomised to one of the three intervention groups, and received their intervention. Two to three months after the intervention the second paper was sent, and approximately six months later the third paper was sent to those who had completed the second review. Here we focus on the difference between the self-taught group and the no-training group in the quality of the review of the second paper.

The trial randomised 173 patients to the no-training group, and 166 to the self-taught group. Of these, 162 and 120 respectively returned a review of the second paper. Our illustrative substantive analysis is the regression of the RQI from the second paper on intervention group and baseline RQI. Using records with no missing data, this found the self-taught package increased the mean RQI by 0.24 units on the RQI scale, (95% CI 0.10–0.38, $p$=0.001).

However, for this substantive analysis, the second paper RQI was missing for 11/173 in the no-training group and 46/166 in the self-taught group. For the substantive analysis, our primary missing data assumption is that data are Missing At Random (MAR).[1] As we elaborate below, this means that — given intervention group and baseline RQI — the distribution of the second paper RQI is the same, whether or not it is observed. This is illustrated in Figure 1, where the fitted lines from the substantive analysis, a regression second paper RQI on baseline and intervention, are shown. Analysis under MAR can be performed using Multiple Imputation (MI)[2], which imputes the missing values from the distribution centred on the fitted lines.

[1] Cross ref: Missing Data
[2] Cross ref: Multiple Imputation


Figure 1: Scatter plot of second paper RQI against baseline paper RQI, showing fitted line assuming RQI is MAR given baseline and intervention. Red triangles show baseline RQI of those with missing second paper RQI.

However, now consider Table 1. This confirms the impression given in Figure 1 that the missing second paper RQI are predominantly from reviewers with a *worse* baseline RQI. It is therefore plausible that these data are missing not at random — that is, their distribution differs in a practically important way from the distribution of the observed data, given baseline and intervention group.

Of course, we cannot determine from the observed data whether this is true or not (in other words whether the data are missing at random or not). Therefore, what we should do is investigate the robustness of our inference about the intervention to contextually plausible assumptions about the missing data. □

| | | No-training group | Self-taught group |
|---|---|---|---|
| Returned review of paper 2 | n | 162 | 120 |
| | mean | 2.65 | 2.80 |
| | SD | 0.81 | 0.62 |
| Did not return review of paper 2 | n | 11 | 46 |
| | mean | 3.02 | 2.55 |
| | SD | 0.50 | 0.75 |

Table 1: For each of the two interventions, this table shows the quality of the review at baseline for (a) those who went on to complete the second review and (b) those who did not.

## 3 Modelling framework for sensitivity analysis

We describe three model classes for data that are Missing Not At Random (MNAR): selection, pattern mixture and latent variable models. All three are very general. However, to keep the exposition accessible, we focus initially on the trial setting above. Extensions to more complex settings build naturally on this.

Suppose our substantive statistical model is a regression of a vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ on covariates $\mathbf{X}$. For our development we suppose $\mathbf{Y}$ is partially observed; however the approach is general and can handle missing covariate

values (Subsection 4.1). Let $R_i = 1$ if $Y_i$ is observed, $R_i = 0$ if it is missing, and let $\mathbf{R} = (R_1, \ldots, R_n)^T$. When data are missing, we need to consider a model for the joint distribution of $\mathbf{Y}, \mathbf{R}$, that is:

$$f(\mathbf{Y}, \mathbf{R}|\mathbf{X}; \boldsymbol{\xi}).$$

We can factor this joint distribution in two ways:

$$\begin{aligned} f(\mathbf{Y}, \mathbf{R}|\mathbf{X}; \boldsymbol{\xi}) &= f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_S)\mathrm{P}(\mathbf{R}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}_S) \quad (1) \\ &= f(\mathbf{Y}|\mathbf{X}, \mathbf{R}; \boldsymbol{\theta}_P)\mathrm{P}(\mathbf{R}|\mathbf{X}, \boldsymbol{\gamma}_P). \quad (2) \end{aligned}$$

The first factorisation, (1) is known as a *selection* model.[3] This partners the substantive scientific model (here the regression of $\mathbf{Y}$ on $\mathbf{X}$, denoted $f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_S)$ with a model for the data 'selected' for observation.

The second factorisation, (2) is know as a *pattern mixture* model.[4] Here, the distribution for $\mathbf{Y}$ explicitly depends on $\mathbf{R}$ (i.e. whether the data are observed), and the overall distribution is a weighted average over these 'patterns'. The weights are the observed frequency of each pattern. Note that in this factorisation the substantive scientific model does not appear explicitly. In principle, the substantive parameters can be derived from $\boldsymbol{\theta}_P, \boldsymbol{\gamma}_P$. In practice, as we shall see below, a much simpler option is to use multiple imputation.

Because the two approaches are different factorisations of the same joint density, each selection model has a corresponding pattern mixture representation and vice versa; however what is simple in one form may be awkward in the other.

The third approach parameterises (1) using latent variables; these can be random coefficients (e.g. intercepts) or latent classes. While this approach has received considerable attention in the literature, it is not so directly amenable to multiple imputation, so we do not consider it further here. For further details see, for example, Wu and Carroll (1988); Wu and Bailey (1998); Kenward and Rosenkranz (2011).

Before moving on, recall that by definition, if values of $\mathbf{Y}$ are MAR [5] (given $\mathbf{X}$) then — once we know $\mathbf{X}$ — $\mathbf{R}$ is independent of $\mathbf{Y}$. In other words, given $\mathbf{X}_i$, the chance of observing $Y_i$ is independent of the actual value of $Y_i$, so that $\mathrm{P}(\mathbf{R}|\mathbf{X}, \mathbf{Y}; \boldsymbol{\gamma}_S) = \mathrm{P}(\mathbf{R}|\mathbf{X}; \boldsymbol{\gamma}_S)$. When this is true in (1), it follows that in (2) $f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_S) = f(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_P)$, i.e. that the distribution of $\mathbf{Y}$ given $\mathbf{X}$ is the same *whether or not* $\mathbf{Y}$ *is observed*.

This key observation is exploited to impute missing values under the missing at random assumption[6]; departing from this means data are MNAR, and so provides a natural approach to sensitivity analysis.

EXAMPLE: *peer review trial (ctd)*

Looking at Figure 1, the preceding paragraphs imply that, assuming RQI is missing at random given fully observed baseline and intervention group, the distribution of the missing RQI values (red triangles) should be the same as the observed ones (blue circles). Multiple imputation under MAR uses this directly to impute the missing values; Figure 2 shows an imputed dataset.

Using the standard multiple imputation procedure, we impute $K = 20$ data sets, fit the substantive model to each, and combine the results using Rubin's rules[7]. The results are shown in the first two rows column of Table 5. In line with theory, the results of multiple imputation under MAR agree closely with fitting the substantive analysis to the observed data. This is because when we have missing data in the dependent variable, multiple imputation cannot recover any missing information from that individual (unit) unless we have strong auxiliary variables[8]. In turn, this is because the contribution to the likelihood for each individual with missing response is 1, as

$$L_i(\boldsymbol{\theta}) = \int f(Y_i|\mathbf{X}_i; \boldsymbol{\theta})\, dY_i = 1.$$

[3] Cross ref: Selection Model (Missing Data)
[4] Cross ref: Pattern-Mixture Model
[5] Cross ref: Missing Data
[6] Cross ref: Multiple Imputation
[7] Cross ref: Multiple Imputation
[8] Cross ref: Multiple Imputation


Figure 2: Scatter plot of second paper RQI against baseline paper RQI, showing a single imputed dataset (imputed values are red triangles).

## 4 Pattern mixture modelling using multiple imputation

As the discussion above shows, once data are missing not at random, the distribution $f(\mathbf{Y}|\mathbf{X})$ will differ according to the missingness pattern $\mathbf{R}$. This complicates the analysis considerably, because

1. these differences can take any form, e.g. mean, variance, skewness etc., and
2. there is no information in the observed data about these differences!

Given this, a practical way forward using multiple imputation is to:

1. start with the imputed values under MAR (recall MAR assumes that $f(\mathbf{Y}|\mathbf{X})$ is the same for missing and observed values of $\mathbf{Y}$), then
2. alter them in (i) the simplest way posible to represent plausible departures from MAR that (ii) are likely to impact on inferences and (iii) are accessible to subject experts.

Points 2(i), 2(ii) and 2(iii) are important. By making the changes as simple as possible, we limit the number of unknown parameters describing the changes, which are known as sensitivity parameters. By focusing on changes that are likely to impact inferences, we focus our efforts where it matters. Finally, by being accessible to subject experts, we make the results interpretable.

### *Generic Algorithm*

We now describe the generic algorithm for sensitivity analysis using pattern mixture models with multiple imputation. To do this, we introduce a more generic notation, showing how this incorporates the above discussion.

For the $i$th of the $n$ units (e.g. individuals), denote the data by $\boldsymbol{Z}_i = (Z_{i,1}, \ldots, Z_{i,p})'$ and let $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,p})'$ be the vector of response indicators. Suppose that there are, across all $n$ units, $M \ll n$ distinct response patterns, indexed by $\mathbf{R}_m, m \in (1, \ldots, M)$. The missing and observed variables from any unit conform to one of these patterns, say $m(i)$, for the $i$th unit, and one of the patterns corresponds to complete records on all $p$ variables. Let $\boldsymbol{Z}^{obs}_{m(i)}, \boldsymbol{Z}^{miss}_{m(i)}$ be the observed and missing variables for unit $i$ with response pattern $m(i)$.

For each unit, we denote the distribution of the missing data, given the observed data and missingness pattern, by

$$f\left(\boldsymbol{Z}_{m(i)}^{miss} \mid \boldsymbol{Z}_{m(i)}^{obs}, m(i), \boldsymbol{\theta}_{m(i)}\right). \tag{3}$$

Here $\boldsymbol{\theta}_m$ are the parameters of this distribution for missingness pattern $m$, whose values we have to estimate before we can draw missing data from (3).

As discussed above, if the missing data are MAR, then (3) does not depend on the missingness pattern $m$; distributions are the same across all patterns so that $\boldsymbol{\theta}_m = \boldsymbol{\theta}$, $m \in (1, \ldots, M)$, and we impute missing data from $f(\boldsymbol{Z}_i^{\text{miss}}|\boldsymbol{Z}_i^{obs}, \boldsymbol{\theta})$. However, if the data are Not Missing at Random (NMAR), then this distribution will differ across missingness patterns.

EXAMPLE: *peer review trial (ctd)*

To relate the above notation to the peer review trial, recall that for each reviewer, the observed data is the baseline review quality, the intervention group, and (where observed) the second paper review quality index. This is therefore $\boldsymbol{Z}_i^{obs}$. For each reviewer, the response indicator is a scalar, $R_i$, which is 1 if the second paper review quality index is observed. When it is missing, its value is denoted by $Z_i^{miss}$.

Thus, the general expression (3) above corresponds to the linear regression model for partially observed paper 2 RQI on baseline and intervention. Assuming MAR, $\boldsymbol{\theta}_m = \boldsymbol{\theta}$, which are the intercept, coefficients of baseline and intervention, and residual variance of the model (Figure 2). □

Suppose that $\hat{\beta} = \hat{\beta}(\boldsymbol{Z}) = \hat{\beta}(\boldsymbol{Z}^{miss}, \boldsymbol{Z}^{obs})$ is of interest. For example, in the peer review study $\beta$ is the intervention effect. With no missing data, we would estimate this from $\boldsymbol{Z}$, using the appropriate regression model. For each missingness pattern $m$, our approach is to define a form for (3) which reflects contextually relevant assumptions. Then we impute $K$ 'complete' datasets under our MNAR assumptions following the standard MI procedure:[9]

MI1: for each pattern, $m$, take a draw from the Bayesian posterior distribution $f(\boldsymbol{\theta}_m|\boldsymbol{Z}^{obs}, m)$ and then

MI2: impute the missing data from (3) using the above draws $\{\boldsymbol{\theta}_m\}_{m=1}^{M}$.

Both steps are repeated to create each imputed dataset. The parameter of interest is then estimated from each imputed data set in turn to give $\hat{\beta}_k$, with standard error $\hat{\sigma}_k$, $k = 1, \ldots, K$. These are then combined using Rubin's rules.[10]

To implement MI1 we need to choose a model for the observed data. To implement MI2 we need to specify (3). Taking the former first, the approach we adopt is to assume MAR and estimate (common across all patterns) $\boldsymbol{\theta}$ from all the observed data. Then, for each pattern $m = 1, \ldots, M$, specific rules or information are used to derive, or draw, $\boldsymbol{\theta}_m$. This is of necessity context specific. It could involve:

1. explicitly specifying the distribution of $\boldsymbol{\theta}_m$ given $\boldsymbol{\theta}$, possibly using opinions elicited from experts, or
2. specifying how $\boldsymbol{\theta}_m$ is constructed from $\boldsymbol{\theta}$, for example in terms of rules across well defined subsets of the data (such as treatment or exposure groups).

This is deliberately framed generally because the way we implement these steps is going to vary by subject area and analysis. To show how this may be done, we now illustrate their application in two examples. The first is the peer review study, which has missing values in the outcome (dependent) variable. The second is a cohort study, with missing values in a covariate and the outcome.

EXAMPLE: *peer review trial (ctd)*

Recall that our substantive analysis is the regression of RQI from the second paper on baseline and intervention, and that — under MAR — we use the parameter estimates from fitting this model to the observed data to impute the missing values (Figure 2). We are now going to explore the robustness of our inferences to departures from MAR, that is we are going to change the imputation distribution. We seek to do this consistent with the three principles discussed above.

[9] Cross ref: Multiple Imputation

[10] Cross ref: Multiple Imputation


Figure 3: Sensitivity analysis using the '$\delta$-method': scatter plot of second paper RQI against baseline paper RQI, showing a single imputed dataset under missing not at random. Imputed values (red triangles) are moved an average of $(\delta_N, \delta_S)$ from MAR.

The focus of the analysis is on the mean difference between the (baseline adjusted) RQI scores by intervention. This suggests sensitivity analysis could usefully focus on changing the mean of the missing RQI scores (conditional on baseline) from their values under MAR, leaving other aspects of the distribution unchanged.

Suppose that this difference is chosen to be $\delta_N$ for the no-training group and $\delta_S$ for the self-taught group. Then the proposal is summarised in Figure 3. The imputation distribution is shifted by $(\delta_N, \delta_C)$ respectively for the two arms, and its mean (given baseline) is shown by the red line. The red triangles show a single imputed data set from this distribution.

This approach follows our principles set out above: it is described by a small number of parameters; it focuses on changes from MAR that directly impact on conclusions, and it can be accessibly presented to experts.

Of course, $(\delta_N, \delta_C)$ cannot be estimated from the data; information about these must be supplied by the analyst. For now we choose two values: $\delta_N = -0.4$, and $\delta_S = -0.6$ points on the RQI scale. This reflects our intuition (Table 1) that reviewers with missing data in the self-taught group are providing poorer reviews.

Following the generic approach set out above, this is straightforward to implement. We simply take the imputed values under MAR, and add $\delta_N$ to those in the no-intervention group, and $\delta_S$ to those in the self-taught group, fit the substantive model to each imputed data set and apply Rubin's rules. The results are shown in Table 5. We see our conclusions are highly sensitive to MAR; if our assumed values of $(\delta_N, \delta_S)$ are correct, then there is no effect if the self-taught package. □

The challenge with this approach is to specify values of the sensitivity parameters, $\boldsymbol{\delta}$. One way is to keep increasing them from zero (which corresponds to MAR) until our conclusions change. We can then step back and decide if those values that change the conclusions are contextually plausible. This is often referred to as a 'tipping-point' analysis. The difficulty is that in general there will be more than one sensitivity parameter (as in the peer review study above) and therefore there will in general be many values that 'tip' the results. This may be challenging to convey to subject experts, in which case it is hard to establish the contextual plausibility of the 'tipping-points'.

Broadly speaking, there are two ways we can address this: first by eliciting the beliefs of experts about the sensitivity parameters; alternatively by framing the sensitivity analysis by reference to other groups of patients. Both can be readily implemented using multiple imputation, and we discuss them below.

### *4.1 Missing covariates*

We now explore how the generic framework introduced above can be applied when we have missing data in a number of variables, and in particular in the covariates.

The presentation above highlights how, even in simple situations, the number of sensitivity parameters quickly increases. Even in the peer review example, we could build on the analysis above, allowing (i) the effect of baseline to vary between groups, (ii) the residual variance to vary between groups, and so on. As having a large number of sensitivity parameters makes it difficult to choose, and justify, their values, it is preferable to try and avoid this by focusing on key missing data patterns, and then exploring them in turn.

For example, consider an observational analysis of an outcome on an exposure, adjusted for confounders. As discussed above, in the absence of strong auxiliary variables, individuals with missing outcome contribute no information. Suppose, as will typically be the case, that the exposure is well recorded, but there are missing values in the confounders. Our focus is on the sensitivity of the exposure effect to the missing at random assumption. To explore this we will want to vary the distribution of the missing confounder(s) given exposure and outcome (see, e.g. Carpenter and Kenward (2013) p. 24–35 and Bartlett *et al.* (2015)). Rather than do this for all the confounders simultaneously, it is usually more accessible to do it for each in turn and then summarise the results.

Suppose the substantive model is a regression of outcome vector $\mathbf{Y}$ on exposure vector $\mathbf{X}$ adjusting for confounders $\mathbf{Z}_1, \mathbf{Z}_2$, and there is a non-trivial proportion of missing data in $\mathbf{Z}_1$. We assume we have multiply imputed $k = 1, \ldots, K$ datasets under MAR. Then, *within each imputed dataset* we proceed as follows:

1. Use an appropriate generalised linear model to regress $\mathbf{Z}_1$ on $\mathbf{Y}, \mathbf{X}, \mathbf{Z}_2$, obtaining coefficients $\hat{\gamma}_0, \ldots, \hat{\gamma}_3$. $\hat{\gamma}_0$ is then the estimated intercept of the linear predictor for $\mathbf{Z}_1$ given values of the other variables.
2. Change $\hat{\gamma}_0$ to $\hat{\gamma}_0 + \delta$, where the user specifies $\delta$, which is the difference—conditional on the other variables—in the intercept of the linear predictor for $\mathbf{Z}_1$ when it is observed/missing.
3. Leave the imputed values of other variables unchanged, but re-impute (using $\hat{\gamma}_0 + \delta$) the missing $\mathbf{Z}_1$ values.

This gives us $K$ imputed datasets under MNAR. Fit the substantive model to each imputed dataset, and combine the results using Rubin's rules. Before implementing this approach, it may be useful to centre $\mathbf{Y}, \mathbf{X}, \mathbf{Z}_2$, so that the intercept, $\gamma_0$, is readily interpretable. This helps identify plausible $\delta$ values.

Note that because Steps 1–3 are applied to each imputed dataset in turn, this preserves the between imputation variation and we do not need to sample from the distribution of the parameters $\hat{\boldsymbol{\gamma}}$ before imputing the missing $\mathbf{Z}_1$ values in Step 3.

We illustrate this approach below. However, while it is very practical, in general it is only approximate. This is because if there are a non-trivial number of missing values in other variables, and we change the imputation distribution for $\mathbf{Z}_1$, then this will in turn affect the values imputed for those other variables. If the multiple imputation has been carried out using the full conditional specification algorithm,[11] it can be adapted address this (Tompsett *et al.*, 2018).

EXAMPLE: *youth cohort study*

To illustrate this approach, we analyse data from the Youth Cohort Time Series for England, Wales and Scotland, which is freely available from the UK data archive, study number SN 5765. We consider data from pupils attending comprehensive schools from five cohorts; these pupils reached the end of Year 11 in 1990, 1993, 1995, 1997 and 1999. We explore the relationship between Year 11 educational achievement, social stratification and ethnicity, using simplified variables summarised in Table 2.

There are 63,629 records, of which 55,222 (87%) are complete, 7,323 are missing `working` alone, 689 are missing `score` alone and 395 are missing both `working` and `score`. The other variables are complete. The probability of missing `working` increases strongly with `non-white` and decreases strongly with increasing `score`.

[11] Cross ref: Multiple Imputation

| Variable name | Description |
|---|---|
| `score` | GCSE points score. Each pupil sits up to 15 GCSE exams. The results for each are converted into a score from 7 (highest grade) to 0 (fail). These are summed across a pupils exams and capped at 84 (equivalent to 12 GCSEs at the top grade). |
| `cohort` | Year of data collection: 1990, 93, 95, 97, 99. |
| `boy` | Indicator for boys. |
| P`working` | 0: Parental occupation is categorised as managerial or intermediate; 1: parental occupation categorised as working. |
| `non-white` | Indicator for non-white ethnicity. |

Table 2: Description of variables in the Youth Cohort Study.

Our substantive analysis is a linear regression of `score` on the other variables; the results of a complete records analysis ($n = 55,222$) are shown in the second column of Table 3. We see that only `non-white` is non-significant, with the results suggesting that children with parents from poorer backgrounds do worse, as do boys, and that exam scores are rising over time. However, since the probability of `working` being missing is much higher for those with a low score and non-white ethnicity, the coefficient of `non-white` is likely to be an underestimate.

Just how much of an underestimate is clear from the second column of Table 3, which show the results from multiple imputation assuming missing at random — that is, assuming that given the other variables, the probability that `working`=1 is the same when it is observed and unobserved. While most of the coefficients are little changed, that for non-white is now highly statistically significant: after adjustment, this group have an average score that is 2.6 points lower.

We now use the algorithm above to explore the sensitivity of the results to the missing at random assumption. We focus on the imputation model for `working`, given the remaining variables. Following the generic algorithm above, we change the adjusted odds of imputing `working`= 1 from those predicted by the observed data, as follows. For each imputed data set, $k = 1, ..., 5$ :

1. Fit the logistic regression

$$\begin{aligned}\text{logit}\{\Pr(\texttt{working}_{i,k} = 1)\} = \gamma_0 + \gamma_1 \texttt{score}_{i,k} + \gamma_2 \texttt{boy}_{i,k} + \gamma_3 \texttt{non-white}_{i,k} \\ + \gamma_4 \texttt{cohort93}_{i,k} + \gamma_5 \texttt{cohort95}_{i,k} + \gamma_6 \texttt{cohort97}_{i,k} + \gamma_7 \texttt{cohort99}_{i,k}\end{aligned}$$

2. Let $\delta$ describe the difference between the adjusted log-odds of `working`= 1 between the observed and the missing data. For each record, $i$ with missing `working` in the original data,

   (a) calculate

$$\begin{aligned}\hat{\eta}_{i,k} = (\delta + \hat{\gamma}_0) + \hat{\gamma}_1 \texttt{score}_{i,k} + \hat{\gamma}_2 \texttt{boy}_{i,k} + \hat{\gamma}_3 \texttt{non-white}_{i,k} \\ + \hat{\gamma}_4 \texttt{cohort93}_{i,k} + \hat{\gamma}_5 \texttt{cohort95}_{i,k} + \hat{\gamma}_6 \texttt{cohort97}_{i,k} + \hat{\gamma}_7 \texttt{cohort99}_{i,k}\end{aligned}$$

   (b) let $\pi_{i,k} = e^{\hat{\eta}_{i,k}}/(1+e^{\hat{\eta}_{i,k}})$, and — for each $i$ with $\texttt{working}_i$ missing — draw $u_{i,k} \sim \text{uniform}[0,1]$ and impute the missing $\texttt{working}_{i,k}$ as 1 if $u_{i,k} < \pi_{i,k}$ and 0 otherwise.

This gives us $K$ imputed datasets, where the missing `working` values have been imputed with an adjusted log-odds that is $\delta$ different from missing at random ($\delta = 0$ corresponds to missing at random).

Then — as usual with multiple imputation — we fit the substantive model to each imputed data set in turn and combine the results for final inference using Rubin's rules.

Applying this algorithm gives the results shown in the rightmost two columns of Table 3. The intercept in the adjusted model for `working`, $\gamma_0$, is $\sim 0.26$, so we explore $\delta = \mp 0.26$. When $\delta = -0.26$, the odds of `working`= 1 are reduced relative to MAR; because of the lower `score` in the non-white and working group the magnitude of the

| Variable | Estimated coefficients for | | | |
|---|---|---|---|---|
| | complete records | missing at random | missing not at random | |
| | | | $\delta = -0.26$ | $\delta = 0.26$ |
| non-white | −0.33 | −2.59 | −2.78 | −2.39 |
| | (0.251) | (0.219) | (0.213) | (0.214) |
| working | −10.26 | −10.63 | −10.40 | −10.88 |
| | (0.154) | (0.173) | (0.169) | (0.157) |
| boy | −3.29 | −3.22 | −3.21 | −3.20 |
| | (0.135) | (0.129) | (0.129) | (0.130) |
| cohort 93 | 5.76 | 5.59 | 5.59 | 5.58 |
| | (0.209) | (0.198) | (0.199) | (0.199) |
| cohort 95 | 9.55 | 9.36 | 9.38 | 9.36 |
| | (0.215) | (0.207) | (0.207) | (0.207) |
| cohort 97 | 8.16 | 8.15 | 8.17 | 8.13 |
| | (0.215) | (0.206) | (0.208) | (0.207) |
| cohort 99 | 12.98 | 12.28 | 13.29 | 13.25 |
| | (0.223) | (0.215) | (0.216) | (0.211) |
| intercept | 36.06 | 35.26 | 35.14 | 35.38 |
| | (0.173) | (0.168) | (0.169) | (0.167) |
| Marginal odds ratio of `working`/ `non-white` | 1.53 | 1.69 | 1.58 | 1.79 |

Table 3: Analyses (estimate, (SE)) of the Youth Cohort Study under various assumptions, showing (bottom row) the marginal odds ratio of non-white ethnicity by 'working' parental occupation category.

`non-white` coefficient increases. Conversely, when $\delta = 0.26$, the magnitude of the `non-white` coefficient decreases. However, the effect is small, so we conclude that the difference between white and non-white ethnicity is robust to the MAR assumption. For a more detailed analysis of these data, see Carpenter and Kenward (2013), p. 240–242. □

### *4.2 Using expert opinion to inform sensitivity parameters*

The above has shown how to use the pattern mixture approach to perform sensitivity parameters for one variable at a time. To do this we need to choose the parameter(s) (referred to as '$\delta$' above) describing the difference between the MAR and MNAR distributions.

An attractive option is to elicit expert opinion on these. This can be viewed as attempting to quantify how experts would implicitly adjust their interpretation of the study due to the missing data. For recent work in this area in observational and trial data see Smuk *et al.* (2017); Mason *et al.* (2017).

EXAMPLE: *peer review trial (ctd)*

Returning to the peer review study, White *et al.* (2007) devised a questionnaire, designed to elicit experts' prior belief about the difference, $\delta$, between the average missing and average observed review quality index in this study. This was completed by 2 investigators and 20 editors and staff at the *British Medical Journal*. The resulting distribution is negatively skewed, with mean $-0.21$ and SD 0.46. Unfortunately, it was not possible to collect information about how this was influenced by the randomised intervention (i.e. whether $\delta_N$ and $\delta_S$ have the same mean, and what their correlation is).

We adopt a bivariate normal model approximation to the prior:

$$\begin{pmatrix}\delta_N \\ \delta_S\end{pmatrix} \sim N\left[\begin{pmatrix}-0.21 \\ -0.21\end{pmatrix}, 0.46^2 \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix}\right]. \tag{4}$$

Because it was not possible to elicit a prior on $\rho$ from the experts, we analyse the data with $\rho = 0$; assuming $\rho \geq 0$, this choice gives the largest standard error for the intervention effect.

Given a draw $(\delta_N, \delta_S)$ from this distribution the pattern mixture model is

$$\begin{aligned}
&\text{RQI}_i = \beta_0 + \beta_1 \text{self-taught}_i + \beta_2 \text{baseline RQI}_i + e_i && \text{if RQI}_i \text{ observed,} \\
&\text{RQI}_i = (\beta_0 + \delta_N) + (\beta_1 + \delta_S - \delta_N)\text{self-taught}_i + \beta_2 \text{baseline RQI}_i + e_i && \text{if RQI}_i \text{ missing,} \\
&\quad e_i \overset{i.i.d.}{\sim} N(0, \sigma^2).
\end{aligned} \tag{5}$$

Thus the mean review quality, relative to that in the observed data, is changed by $\delta_N$ in the no intervention arm and $\delta_S$ in the self-taught arm.

Following the general approach for estimating pattern mixture models via MI described above, we can proceed as follows:

1. Fit the imputation model to the observed data. For $k = 1, \ldots, K$, draw from the posterior distribution of the imputation model parameters $\boldsymbol{\theta}^k = (\beta_0^k, \beta_1^k, \beta_2^k, \sigma^{2,k})$.
2. Draw $(\delta_N^k, \delta_S^k)$ from (4)
3. Using the draws obtained in steps 1 and 2, impute the missing $\text{RQI}_i^k$ using (5).

Steps 1–3 are repeated to create $K = 20$ imputed data sets. Then we fit the substantive model to each imputed data set and apply Rubin's rules.

The results using this expert prior are shown in Table 5. The point estimate is reduced relative to the MAR analysis, but not nearly so much as with the fixed-delta method. The big difference is that the standard error has increased by about 2.5 times, reflecting the variability between the experts. The consequence is the results are no longer statistically significant. □

| Name | Description |
|---|---|
| Reference-based multiple imputation methods: | |
| Jump to reference (J2R) | Imputes assuming that following dropout a patients mean profile follows that observed in the reference arm. Pre-drop out means come from the randomised arm. |
| Copy reference (CR) | The conditional profile given the history is copied from the reference group i.e. imputes as if randomised to reference arm, pre- and post-drop out means come from the reference arm. |
| Last mean carried forward (LMCF) | Forms post-dropout means by carrying forward the randomised arm mean at dropout. |
| Copy increments in reference (CIR) | Forms post-dropout means by copying increments in the reference arm. Pre-drop out means come from the randomised arm. |
| External information multiple imputation methods: | |
| The $\delta$-method | Impute under randomised arm MAR and subtract/add by fixed $\delta$. |

Table 4: Examples of reference-based and external information MNAR multiple imputation methods.

### *4.3 Information anchored sensitivity analysis*

As illustrated above, expert prior elicitation can often result in marked loss of information when we move from MAR to MNAR. The peer review study also illustrated another difficulty of prior elicitation, and sensitivity analysis more generally. This is that even simple models give rise to more sensitivity parameters than can be meaningfully — and certainly readily — elicited from experts.

In response to this, in the context of continuous data, Carpenter *et al.* (2013) propose that, under MNAR, the imputation distribution could be derived *by reference* to other trial arms, or other groups of patients. For instance, when a patient withdraws from the active arm, we could impute them as if they 'jumped to reference'. Or, if they were randomised to an active arm but never received/took the active intervention, we can impute them as if they belonged to the control arm. As Carpenter *et al.* (2013) describe, many other options are possible, making this approach is very flexible, while at the same time avoiding explicit specification of the sensitivity parameters.

Further, when implemented by multiple imputation, Cro *et al.* (2018) show that the results are *information anchored*. This means that the proportion of (statistical) information lost due to missing data when the analysis is carried out assuming missing at random is the same as the proportion lost under a reference based sensitivity analyses. Some common *reference based* analyses are outlined in Table 4; note that the 'delta' method, introduced above, also has the information anchoring property. All these are examples of what is often referred to as *controlled imputation* methods, in other words imputation methods where the analyst directly *controls* some, or all, of the imputation model parameters.

Reference based senstitity analysis for continuous data is implemented in the Stata program `mimix` (Cro *et al.*, 2016), and also in SAS macros available from `www.missingdata.org.uk`. The ideas have recently been extended to the survival setting (Atkinson *et al.*, 2018).

Example: *peer review trial (ctd)*

We illustrate reference based sensitivity analysis using the reviewer trial. In this trial, it is plausible that reviewers randomised to the self-taught group, who did not return the review of paper 2, did not meaningfully engage with the study after completing the review of paper 1. Therefore, for our sensitivity analysis we can impute them as if they were randomised to the no-intervention, which is the *reference* group (cf Table 4).

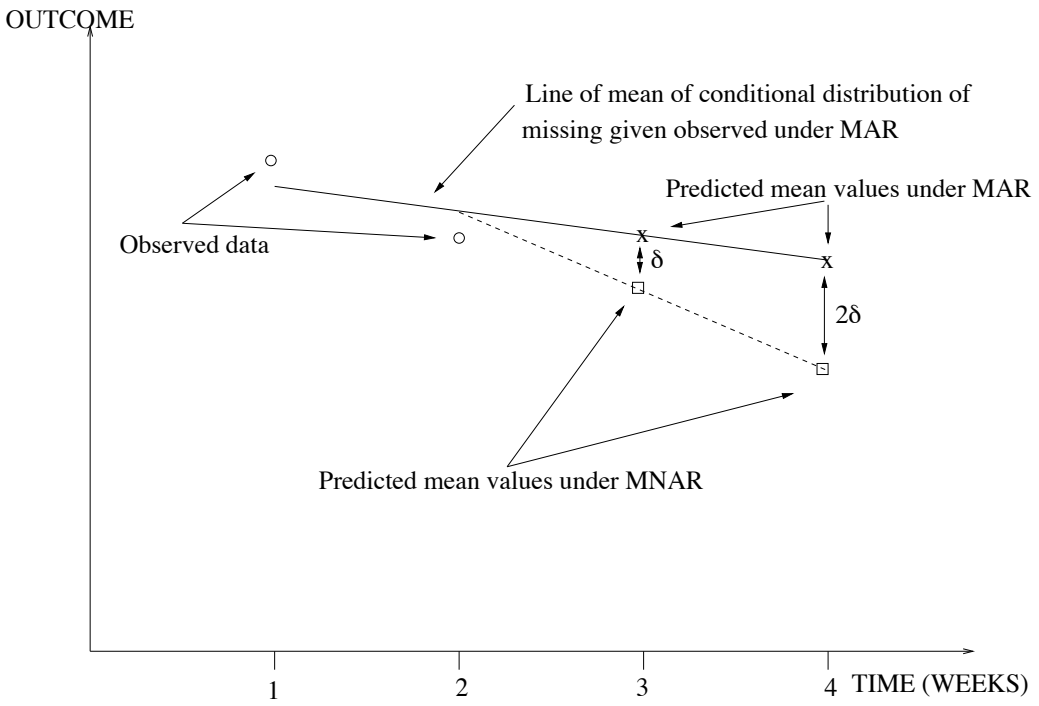


Figure 4: Schematic illustration of increasing the rate of decline by $\delta$ post withdrawal.

The results of doing this are show in Table 5. We see that the point estimate (0.17) is lower than that obtained using the experts prior (0.21). This is because the reviewers in the self-taught arm whose paper 2 RQI scores are missing predominantly produced lower quality reviews of paper 1 (Figure 1). Imputing their missing values conditional on these as if they are in the self-taught group notably reduces the point estimate.

However, comparing the standard errors from the two analyses, the expert prior gives 0.178, while the reference based approach is 0.069. In other words, compared to the primary analysis under MAR and the reference based analysis (which loses the same fraction of information due to missing data), the expert prior loses approximately an additional

$$\left[\frac{\frac{1}{0.069^2}-\frac{1}{0.178^2}}{\frac{1}{0.069^2}}\right]\times 100 = 85\%$$

of the information, because of the heterogeneity across experts. This additional loss of information may or may not be reasonable, but it is good to be aware of it. □

### *4.4 Longitudinal studies*

The pattern mixture approach naturally extends to longitudinal follow-up with missing values. Once again, we assume the missing data assumption for the primary analysis is missing at random. We proceed by calculating the predictive distribution for a patient who withdraws — conditional on their pre-withdrawal data — and then, prior to imputation, changing this to reflect plausible departures from missing at random.

The difficulty is that general changes involve choosing values for a large number of parameters, which makes the results difficult to interpret. To circumvent this, two approaches may be useful: change of slope on withdrawal, and reference based approaches. We now discuss each of these briefly.

Figure 4 illustrates the 'change of slope after withdrawal' approach. The outcome measure is scheduled to be recorded at each of 4 weeks following randomisation. However, for the individual shown, data are only available at weeks 1 and 2. Assuming data are MAR, the predicted means are shown by '×'. Moving to MNAR, the sensitivity analysis increases the predicted mean by $\delta$ the first time after dropout, $2\delta$ the second time and so on. This approach is therefore a natural extension of the method applied in the reviewer study above. As in that case, if we have

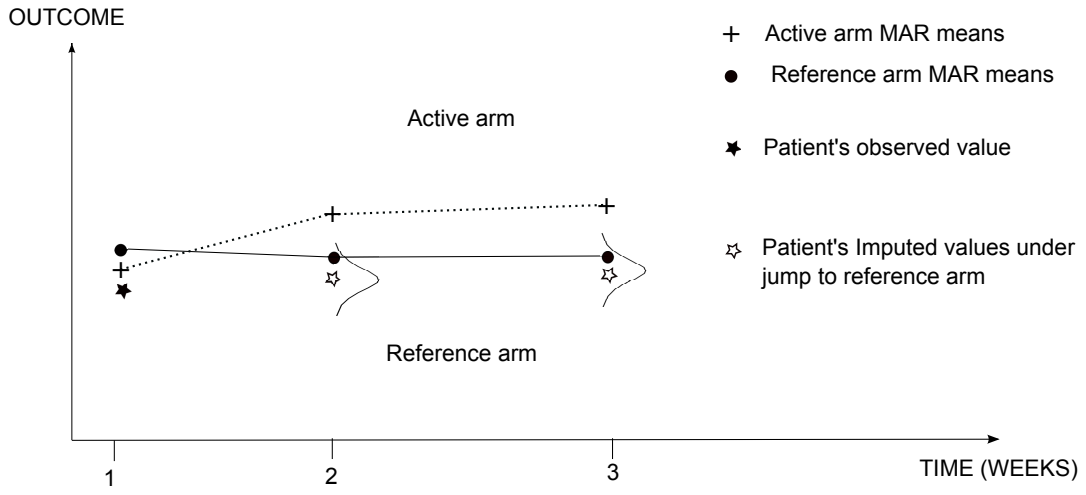


Figure 5: Schematic illustration of *jump-to-reference*.

multiple arms, it may be appropriate to have arm-specific choices for $\delta$. Once $\delta$ is chosen, the data are imputed; then substantive model is fitted to each imputed data set before combining the results for final inference using Rubin's rules.

Figure 5 illustrates the reference-based sensitivity analysis 'jump-to-reference'. The outcome measure is scheduled to be recorded at weeks 1, 2, and 3, for patients from a trial's *active* and *reference* arm. Under the missing at random assumption, the estimated mean at each time is shown for each arm. Now consider an individual randomised to the active arm, whose data are only available at week 1 (show with a '$\star$'). Under jump-to-reference, their post-withdrawal means come from the reference arm, and their covariance crosses over to the reference arm; possible imputed values are shown with a ✡. As usual, once the data are imputed under the assumption, the substantive model is fitted to each imputed data set and the results combined for inference using Rubin's rules.

While, with longitudinal data, there are a limitless number of options for specifying the post-deviation distribution, these information-anchored approaches provide a practical starting point for many applications.

## 5 Approximating selection modelling using multiple imputation

We now outline how multiple imputation may be used to approximate a selection model (1). This approach is particularly useful when the analyst is not the imputer, as instead of creating new imputations (as above), the given imputations are instead re-weighted to approximate a MNAR distribution.

To illustrate, suppose our substantive model is a regression of partially observed $\mathbf{Y}$ on a fully observed $\mathbf{X}_1, \mathbf{X}_2$, that is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2). \tag{6}$$

We wish to explore sensitivity of our results to MNAR assumptions about the partially observed variable $\mathbf{Y}$ with missingness indicator $\mathbf{R}$. Consider the selection model for the probability of observing an individual's data:

$$\text{logit}\{\Pr(R_i = 1)\} = \alpha_0 + \alpha_1 Y_i + \alpha_2 X_{1,i} + \alpha_3 X_{2,i}. \tag{7}$$

Because $R_i = 0$ whenever $Y_i$ is missing, we cannot fit this model alone; we have to fit it jointly with (6). Typically this involves numerical integration to integrate out the missing $Y_i$ and obtain the likelihood which can then be maximised.

However, it turns out that the answer can be readily approximated by first imputing under MAR, and then re-weighting the imputations so that the weighted sample approximately represents the distribution under MNAR. We now give the algorithm; for its justification see Carpenter and Kenward (2013), p. 259–262.

To obtain an estimate of any of the coefficients $\beta_0, \beta_1, \beta_2$ when data are MNAR, the analyst first chooses a plausible value of $\alpha_1$ in (7). Suppose we re-order the data set so that units $i = 1, \ldots, n_1$ have missing $Y_i$. Let $Y_i^k$ denote the $k^{th}$ MAR imputation of $Y_i$. For each imputation, $k$, compute

$$\tilde{w}_k = \exp\left(\sum_{i=1}^{n_1} -\alpha_1 Y_i^k\right), \quad \text{and} \quad w_k = \tilde{w}_k / \sum_{k=1}^{K} \tilde{w}_k. \tag{8}$$

Let $\hat{\beta}_j^k$ $(j = 0, 1, 2, 3)$ denote the estimate of $\beta_j$ from fitting the substantive model to the $k^{th}$ imputed data set (imputed assuming MAR).

Then, under the selection model, (7) with the analyst's choice of $\alpha_1$, the estimate of $\beta_j$ $(j = 0, 1, 2, 3)$ and its variance are

$$\hat{\beta}_{j,\text{MNAR}} = \sum_{k=1}^{K} w_k \hat{\beta}_j^k, \tag{9}$$

with variance

$$\widehat{V}_{\text{MNAR}} \approx \widetilde{V}_W + (1 + 1/K)\widetilde{V}_B, \tag{10}$$

where

$$\widetilde{V}_W = \sum_{k=1}^{K} w_k \hat{\sigma}_k^2, \quad \widetilde{V}_B = \sum_{k=1}^{K} w_k (\hat{\beta}_j^k - \hat{\theta}_{\text{MNAR}})^2. \tag{11}$$

$\hat{\beta}_{j,\text{MNAR}}$ and $\hat{V}_{\text{MNAR}}$ can then be used for inference.

Example: *peer review trial (ctd)*

We use the peer review trial to illustrate this approach. Specifically, we investigate the sensitivity of the results assuming MAR to the possibility that the review quality index from paper 2, $Y_i$, is missing not at random. Let $R_i = 1$ if $Y_i$ is observed and 0 otherwise. We use the following selection model

$$\text{logit}\,\Pr\{(R_i = 1)\} = \alpha_0 + \alpha_1 X_{1,i} + \alpha_2 X_{2,i} + \alpha_3 Y_i \tag{12}$$

where $\mathbf{X}_1$ is the review of the baseline paper and $\mathbf{X}_2$ is an indicator for the self-training group. If $\alpha_3 = 0$, then $Y$ is MAR. As $\alpha_3$ increases from 0 the probability of $Y$ being observed increases with $Y$ (*i.e.* increases with the review quality).

If $\alpha_3 = 0$ (i.e. we omit $\mathbf{Y}$ from (12)), we can fit this model using logistic regression, giving $\hat{\alpha}_1 = 0.21$, $\hat{\alpha}_2 = -1.75$. This shows the probability of withdrawing decreases as baseline review quality increases, and is much higher in the self-training arm.

The estimate $\hat{\alpha}_1 = 0.21$ suggests that each point rise in the baseline average Review Quality Index increases the odds ratio of response by 1.23. In the light of this we carry out sensitivity analyses with $\alpha_3 = 0.3$, which corresponds, on the odds-scale, to roughly a 10% stronger adjusted association between the chance of seeing the second review and its quality.

Using the re-weighting approach (9) and (11) with $K = 150$ imputations gives the results in the bottom row of Table 5. As expected, the estimated effect of the self-training intervention is reduced, but it remains significant at the 5% level. □

### *5.1 Reliability of the approximation*

The weights for this approach are derived by considering the ratio of the desired full-data density $f(\mathbf{Y}|\mathbf{X}, R = 0)$ to the full-data density $f(\mathbf{Y}|\mathbf{X}, R = 1)$. We have to estimate $f(\mathbf{Y}|\mathbf{X}, R = 1)$ from the observed data. However, when the approach is implemented using multiple imputation, the data are drawn from the imputation distribution (i.e. the Bayesian posterior distribution of the missing data given the observed). This takes into account uncertainty in estimating parameters of $f(\mathbf{Y}|\mathbf{X}, R = 1)$, and so the resulting draws are not strictly from the full-data distribution $f(\mathbf{Y}|\mathbf{X}, R = 1)$, but have heavier tails.

| Analysis | Est | SE | p-value | 95% CI |
|---|---|---|---|---|
| Missing at random | | | | |
| Observed data | 0.24 | 0.070 | 0.001 | (0.10, 0.38) |
| MI under MAR, $K = 20$ | 0.23 | 0.07 | 0.002 | (0.09, 0.37) |
| Missing not at random | | | | |
| '$\delta$-method,' $\delta_N = -0.4, \delta_C = -0.6$ : | 0.03 | 0.071 | 0.686 | (-0.11, 0.17) |
| Expert's Prior: | 0.21 | 0.178 | 0.250 | $(-0.158, 0.575)$ |
| Copy no intervention (i.e. reference) | 0.17 | 0.069 | 0.013 | (0.037, 0.307) |
| Selection weighting, $\gamma = 0.3$ (150 imputations) | 0.20 | 0.069 | 0.004 | (0.065, 0.335) |

Table 5: Peer review study: estimated effect of self-taught package assuming data are missing at random, and under various missing not at random analyses.

In applications, where we typically have hundreds of observations and up to around 30% are missing, this is unlikely to cause a problem. However, as discussed in Carpenter *et al.* (2007), and explored further by Rezvan *et al.* (2015) in smaller sample sizes, and particularly with binary data, while the re-weighted estimates are always in the right direction this can cause over-correction. This is because the weights are too large for the tails of the MI draws from $f(\mathbf{Y}|\mathbf{X}, R = 1)$. Smuk (2015) details the reasons for this, proposes a correction and shows this works well for normal data; see also Carpenter and Kenward (2013), p. 262–3. The method is also explored in the context of missing covariates by Bousquet *et al.* (2012).

## 6 Discussion and further reading

When a non-trivial proportion of data are missing, the above examples illustrate that it is practically important to explore the robustness of inferences under the missing at random assumption (which is generally the most natural missing data assumption for the primary analysis) to contextually plausible missing not at random mechanisms.

Hopefully, the examples above illustrate that multiple imputation provides a very flexible and practical approach to exploring the sensitivity of inferences to the inherently untestable assumption that data are missing at random. While the details provided above should allow readers to implement the approaches in more standard settings, for more details and examples readers are referred to Carpenter and Kenward (2015) and Carpenter and Kenward (2013), ch. 10.

## 7 Acknowledgements

James Carpenter is supported by the Medical Research Council, grant numbers MC_UU_12023/21 and MC_UU_12023/29. This article was completed during a research visit to the Malawi Epidemiology and Research Unit.

## Related Wiley StatsRef articles

Bayesian Analysis and Markov Chain Monte Carlo Simulation
Bayesian Methods
Bias from Nonresponse
Clinical Trials and Intervention Studies
Coarse Data
Coarsening at Random
Cohort Studies
Covariates
Data Augmentation
DiggleKenward Model for Dropouts
Dropouts in Longitudinal Studies: Methods of Analysis
Dropouts in Longitudinal Surveys
Elicitation
EM Algorithm
Generalized Linear Models for Longitudinal Data
HorvitzThompson Estimator
Imputation
Incomplete Data
Latent Class Analysis
Loss of Information
Missing Data
Missing Data and Imputation
Missing data estimation, hot deck and cold deck
Missing Data in Clinical Trials
Missing Data in Epidemiologic Studies
Missing Data: Sensitivity Analysis
Multiple Imputation
Multiple Imputation Methods
Nonignorable dropout in Longitudinal studies
Nonresponse in Sample Surveys
OutcomeDependent Selection Models
PatternMixture Model
Sample Size Determination in Clinical Trails Including Adjustments for
Noncompliance, Loss to FollowUp and Groups Sequential Analysis
Selection Model (Missing Data)
Sensitivity Analysis in Observational Studies
Sensitivity Analysis: Overview
Stochastic EM

## References


Atkinson, A., Kenward, M., Clayton, T. and Carpenter, J. R. (2018) Reference based sensitivity analysis for time-to-event data, *Under revision for Pharmaceutical Statistics*.

Bartlett, J. W., Harel, O. and Carpenter, J. R. (2015) Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American Journal of Epidemiology*, **182**, 730–736.

Bousquet, A. H., Desenclos, J. C., Larsen, C., Le Strat, Y. and Carpenter, J. R. (2012) Practical considerations for sensitivity analysis after multiple imputation applied to epidemiological studies with incomplete data. *BMC Methodology*, **12**, 73.

Carpenter, J. R. and Kenward, M. G. (2013) *Multiple Imputation and its Application*. Chichester, Wiley.

Carpenter, J. R. and Kenward, M. G. (2015) Sensitivity analysis with multiple imputation. In *Handbook of missing data methodology* (Eds G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis and G. Verbeke), pp. 435–470. Florida, CRC press.

Carpenter, J. R., Kenward, M. G. and White, I. R. (2007) Sensitivity analysis after multiple imputation under missing at random — a weighting approach. *Statistical Methods in Medical Research*, **16**, 259–275.

Carpenter, J. R., Roger, J. H. and Kenward, M. G. (2013) Analysis of longitudinal trials with protocol deviation:— a framework for relevant accessible assumptions and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, **23**, 1352–1371.

Cro, S., Morris, T. P., Kenward, M. G. and Carpenter, J. R. (2016) Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *Stata Journal*, **16**, 443–463.

Cro, S., Carpenter, J. R. and Kenward, M. G. (2018) Information anchored sensitivity analysis, *revision submitted for publication*.

Kenward, M. G. and Rosenkranz, G. (2011) Joint modelling of outcome, observation time and missingness. *Journal of Biopharmaceutical Statistics*, **21**, 252–262.

Mason, A. J., Gomes, M., Grieve, R., Ulug, P., Powell, J. T. and Carpenter, J. R. (2017) Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the improve trial. *Clinical Trials*, **14**, 357–367.

Rezvan, P. H., White, I. R., Lee, K. J., Carlin, J. B. and Simpson, J. A. (2015) Evaluation of a weighting approach for performing sensitivity analysis after multiple imputation. *BMC Medical Research Methodology*, **15**, 83.

Schroter, S., Black, N., Evans, S., Carpenter, J., Godlee, F. and Smith, R. (2004) Effects of training on quality of peer review: randomised controlled trial. *British Medical Journal*, **328**, 673–675.

Smuk, M., Carpenter, J. R. and Morris, T. P. (2017) What impact do assumptions about missing data have on conclusions? a practical sensitivity analysis for a cancer survival registry. *BMC Med Res Methodol*, **17(1)**, 21.

Smuk, M. J. (2015) *Missing data methodology: sensivitiy analysis after multiple imputation*. Ph.D. thesis, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E7HT.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J. and White, I. R. (2018) On the use of the not-at-random fully conditional specification (narfcs) procedure in practice. *Statistics in Medicine*, **XX**, XX–XX.

van Rooyen, S., Black, N. and Godlee, F. (1999) Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *J Clin Epidemiol*, **52**, 625–9.

White, I., Carpenter, J., Evans, S. and Schroter, S. (2007) Eliciting and using expert opinions about non-response bias in randomised controlled trials. *Clinical Trials*, **4**, 125–139.

Wu, M. C. and Bailey, K. R. (1998) Analysing changes in the presence of informative right censoring by modelling the censoring process. *Statistics in Medicine*, **7**, 175–188.

Wu, M. C. and Carroll, R. J. (1988) Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics*, **44**, 175–188.