Received: 30 July 2019 | Revised: 25 March 2020 | Accepted: 18 April 2020

DOI: 10.1002/sim.8569

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide

Suzie Cro[1] | **Tim P. Morris**[2,3] | **Michael G. Kenward**[4] | **James R. Carpenter**[2,3]

[1]Imperial Clinical Trials Unit, Imperial College London, London, UK

[2]MRC Clinical Trials Unit at UCL, UCL, London, UK

[3]Medical Statistics Department, LSHTM, London, UK

[4]Ashkirk, Scotland, UK

Correspondence
Suzie Cro, Imperial Clinical Trials Unit, Imperial College London, Stadium House, 68 Wood Lane, London W12 7RH, UK.
Email: s.cro@imperial.ac.uk

Missing data due to loss to follow-up or intercurrent events are unintended, but unfortunately inevitable in clinical trials. Since the true values of missing data are never known, it is necessary to assess the impact of untestable and unavoidable assumptions about any unobserved data in sensitivity analysis. This tutorial provides an overview of controlled multiple imputation (MI) techniques and a practical guide to their use for sensitivity analysis of trials with missing continuous outcome data. These include δ- and reference-based MI procedures. In δ-based imputation, an offset term, δ, is typically added to the expected value of the missing data to assess the impact of unobserved participants having a worse or better response than those observed. Reference-based imputation draws imputed values with some reference to observed data in other groups of the trial, typically in other treatment arms. We illustrate the accessibility of these methods using data from a pediatric eczema trial and a chronic headache trial and provide Stata code to facilitate adoption. We discuss issues surrounding the choice of δ in δ-based sensitivity analysis. We also review the debate on variance estimation within reference-based analysis and justify the use of Rubin's variance estimator in this setting, since as we further elaborate on within, it provides *information anchored* inference.

KEYWORDS

clinical trials, controlled multiple imputation, missing data, multiple imputation, sensitivity analysis

1 | INTRODUCTION

In late-phase clinical trials, loss to follow-up and intercurrent events—such as treatment withdrawal or partial compliance—are almost inevitable. Consequently, we often cannot measure what we intended to for all individuals. Planned outcomes may be unobtainable due to the type of the deviation (eg, missed patient visit) and, depending on the nature of the estimand and analysis, even values that were recorded post deviation may be best regarded as missing (eg, data post treatment withdrawal when an on-treatment estimand is of interest). When missing data occurs complexity arises, since any statistical analysis necessarily makes an untestable assumption about the distribution of the unobserved

data. If the wrong assumption is made, the obtained treatment effect and its standard error will be biased, resulting in misleading inferences. To understand how far key inferences depend on the missing data assumption, analysis of incomplete data should therefore consist not only of a primary analysis, under the most plausible missing data assumption, but include sensitivity analyses, which make a range of different credible assumptions for the unobserved data. Sensitivity analysis addresses the same clinical question as the primary analysis, but under contrasting assumptions in order to assess how robust or sensitive results are Reference 1.

Regulatory guidelines from the European Medicines Agency (EMA, 2010)[2] and a Food and Drug Administration (FDA) mandated panel report from the US National Research Council (2010)[3] emphasize the importance of conducting sensitivity analysis in this context. And various methods exist for conducting such sensitivity analyses in the clinical trial arena.[4-6] However, despite these guidelines and methodological developments, recent reviews have highlighted that only around a third of trials with missing data are reporting sensitivity analyses.[7,8] This indicates a large gap between methodological developments and practical application. The more recent publication of the ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials (2019)[9] elaborates further on the importance and provides a framework for how such sensitivity analysis should be approached. Together, these reports highlight the need for accessible and relevant methods of sensitivity analysis, where the changes in assumptions are directly applicable to the primary analysis and can be understood by key stakeholders.

One approach that enables contextually relevant accessible sensitivity analysis of clinical trials with missing data is *controlled multiple imputation* (MI). Controlled MI procedures combine pattern-mixture modeling with MI to provide a practical platform for sensitivity analysis. Controlled MI procedures include δ-based methods, which enable one to explore the impact of a worse or better outcome than that predicted based on the individuals observed data, and the outcomes of similar patients with observed outcome data. An alternative example of controlled MI is reference-based MI, which enables one to explore the impact of individuals with missing data behaving like a specified reference group.

Such controlled MI procedures enable accessible assumptions to be explored to evaluate the impact of missing data. Furthermore, complex lengthy coding can be avoided via MI since standard statistical software packages with inbuilt MI programs can be utilized for analysis. For example, `mi impute` within Stata, `proc mi` within SAS, and the R package `mice`. The recent increase in the discussion of controlled MI methods in the literature[10-16] has drawn attention to their use for contextually relevant accessible sensitivity analysis of longitudinal trials. For examples of their use, see the analyses in References 17-20.

The purpose of this tutorial is to provide an overview of controlled MI procedures for missing data sensitivity analysis and a practical guide to their use for a continuous outcome, with worked examples and Stata code. We demonstrate the applicability and accessibility of the methods for estimating both treatment policy and hypothetical estimands. First, in Section 2, we introduce our two motivating trial case studies, which we will use to demonstrate sensitivity analysis via controlled MI. In Section 3, we discuss estimands and the problem of handling missing data within the analysis of clinical trials in more depth, followed by an outline of our general approach to primary and sensitivity analysis. This includes the necessary background on the MI procedure. We subsequently focus on the two aforementioned controlled MI approaches. The first, δ-based MI, is described and illustrated in Section 4. Issues around the choice of δ are discussed. The second controlled MI approach, reference-based MI, is described and illustrated in Section 5. We demonstrate how different assumptions for unobserved data can be readily made for different groups of individuals in the same trial analysis. In Section 6, we review the debate on variance estimation within reference-based MI analysis and justify the use of Rubin's variance estimator in this setting since, as we later further elaborate on, it provides *information anchored* inference. We end by discussing both methods and their use within clinical trials in Section 7.

2 | MOTIVATING EXAMPLES

2.1 | The ADAPT trial

The Atopic Dermatitis Anti-IgE Paediatric Trial (ADAPT) was a single center, double-blind, randomized controlled trial conducted to determine whether the anti-IgE treatment, omalizumab, improves eczema severity compared to placebo in children. A total of 62 eligible children with severe eczema were randomized to treatment with omalizumab (n=30) or placebo (n=32) for 24 weeks. The trial protocol, statistical analysis plan, and main results have been previously reported.[21-23] In summary, significant and clinically important treatment effects at 24 weeks were reported for eczema severity and measures of quality of life, including the (Children's) Dermatology Life Quality Index Questionnaire or

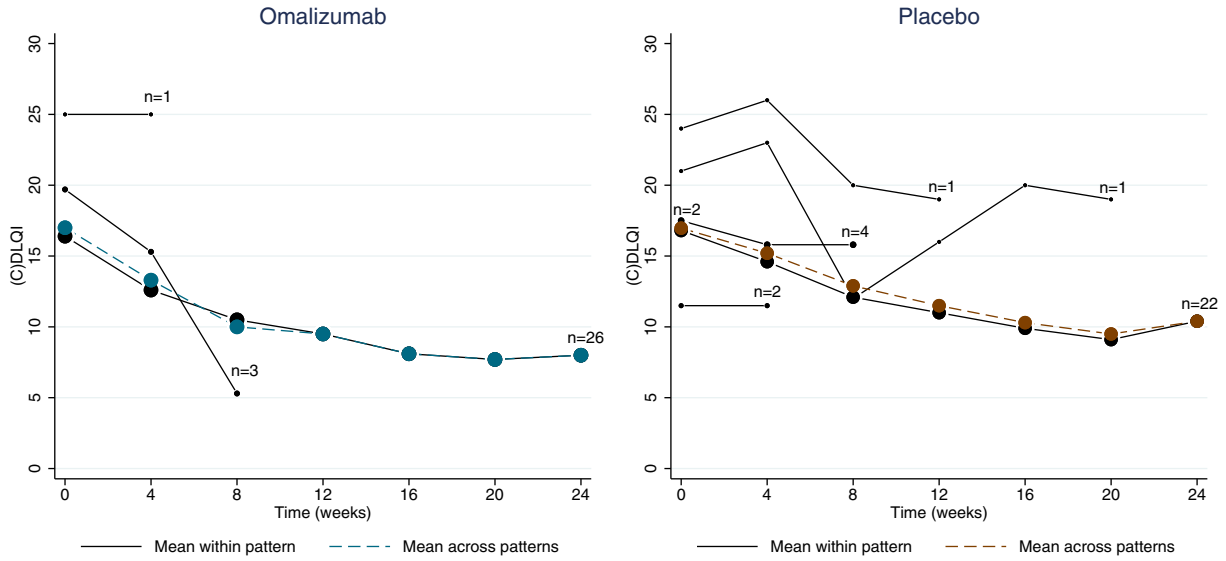

FIGURE 1 Observed mean profiles in the ADAPT trial [Colour figure can be viewed at wileyonlinelibrary.com]

(C)DLQI. The (C)DLQI results in a numerical score ranging between 0 and 30 (higher scores indicating worse quality of life). Data completion rates were high with only 2 participants missing week 24 follow-up for (C)DLQI in the placebo arm, who had previously withdrawn from treatment. However, 11 individuals (7 placebo 4 omalizumab) received rescue medication (alternative systemic therapy or oral steroids) sometime post week 8, up to week 24. An additional placebo patient withdrew from treatment just after week 8. Of note, two of the placebo patients who received rescue medication also withdrew from treatment thus deviated twice from the protocol.

We are interested in the treatment effect on the (C)DLQI in the absence of rescue medication/treatment withdrawal, that is, under on-treatment (hypothetical) conditions. For the purpose of our analysis, data collected post the use of rescue medication/treatment withdrawal will thus be considered missing; it is not relevant to the treatment effect or estimand of interest. Figure 1 shows the missing data patterns by treatment arm and mean (C)DLQI by missing data pattern. There are notably more deviators in the placebo arm than in the active arm.

Under on-treatment conditions, it is most plausible that the unobserved data would have been similar to the data for those with similar characteristics and (C)DLQI profile (up to time of deviation) still in the trial under on-treatment conditions. Or in other words, as we will define formally in Section 3, that the unobserved data can be considered to be missing-at-random. But we will never be certain that this assumption is true and that the results of this analysis are therefore valid. It is conceivable that participants' week-24 responses could have been worse than for those observed with similar characteristics in the absence of rescue mediation. Furthermore, for those who withdrew from treatment, since they decided not to progress, it is also likely that they could have had a worse response than for those observed. In Section 4, we show how δ-based MI can be used to conduct a sensitivity analysis which makes such assumptions.

2.2 | Acupuncture for chronic headaches

Our second example is a multicenter randomized controlled trial conducted by Vickers et al[24,25] of acupuncture for chronic headaches. A total of 401 patients with chronic headache were randomized to standard care or acupuncture and standard care for 12 months. One of the main trial outcomes was a headache score measured at baseline, 3 months and 12 months. However, not all of the randomized patients completed the 12-month follow-up. A total of 44 acupuncture participants and 56 standard care participants withdrew from the trial at or prior to 12 months. The main reasons given for withdrawal included withdrawal of consent, lost to follow-up, or intercurrent illness (see Table 4). Figure 2 shows the missing data patterns within the data. In total, 25% (100/401) of the participants were missing month 12 data.

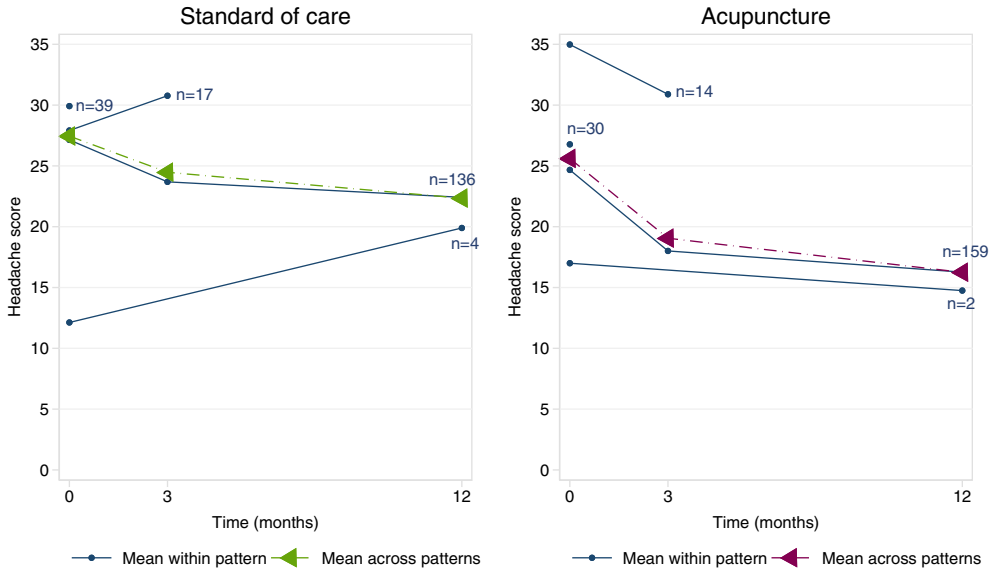

FIGURE 2 Observed mean profiles in the acupuncture trial [Colour figure can be viewed at wileyonlinelibrary.com]

Here, we are interested in the effect of being allocated to acupuncture, that is, a treatment policy estimand, but the analysis is complicated by the unobserved data. Data are not available post trial withdrawal. As in ADAPT, it is most plausible that the unobserved data post trial withdrawal is missing-at-random. But it is also plausible that, following trial withdrawal, patients in the acupuncture arm experienced outcomes similar to those in the standard care arm. Figure 2 shows how on average, headache scores were slightly worse in the standard care arm at all time points. Alternatively, patients who withdrew in the control arm might subsequently have taken up acupuncture. We will explore the impact of patients behaving like those in different arms of the trial on the results in Section 5, using reference based MI. This controlled MI method is appealing since it avoids having to specify any numerical sensitivity analysis parameters; only qualitative assumptions are required.

3 | ANALYSIS WITH MISSING DATA

3.1 | Estimands

In any trial, such as ADAPT or the acupuncture trial, we can only begin to think about missing data once we know the precise treatment effect we are aiming to estimate. Generically, the term *estimand* refers to what is being estimated. Within the clinical trial context, as described in the ICH E9 addendum, an estimand refers to the precise definition of the treatment effect to be estimated to address the scientific question of interest posed by the trial's primary objective. It should capture exactly for whom, what, and when the trials intervention effect to be estimated applies, to meet the clinical goals of the trial and analysis. ICH E9 recognizes five key attributes that, when fully specified together, form a description of an estimand. These are:

(A) The population; the patients targeted by the scientific question,
(B) The treatment condition of interest and the alternative treatment condition(s), for example, control or placebo,
(C) The variables (or endpoint) to be obtained for each patient required to address the scientific question,
(D) The specification of how to account for intercurrent events to reflect the scientific question of interest,
(E) The population level summary for the variable, which provides a basis for a comparison between treatment conditions.

Specification of attribute (D) is critical. For example, do we want to estimate the effect of an intervention regardless of the occurrence of intercurrent events such as treatment withdrawal or receipt of rescue medication. Such an analysis

strategy has been termed "treatment policy." This type of estimand is of interest for the acupuncture trial. Or do we alternatively want to estimate the effect of an intervention under hypothetical conditions, such as under ideal on-treatment conditions only. This latter hypothetical estimand is of interest in our analysis of the ADAPT trial.

The design of a trial and data collection should be aligned with the choice of estimand. For example, if we are interested in a treatment-policy estimand, we will want to ensure data collection following the occurrence of intercurrent events, such as after treatment withdrawal or use of rescue medication. If we are interested only in estimating the treatment effect under on-treatment conditions, the time and cost of data collection should be weighed against the benefit of collecting such data. If an estimand of the latter type is of interest, even if data post treatment withdrawal is collected (as in the ADAPT trial), this may be best set to missing for the purpose of analysis. Data under on-treatment conditions do not exist for the patients who withdrew from treatment or received rescue therapy following their time of withdrawal/additional treatment receipt.

The key point to re-emphasize is that we can only begin to think about missing data in any trial setting once we have fully specified the precise treatment effect we wish to estimate. The estimand should inform what data is missing for the analysis and how missing data should be handled in the analysis. Failure to collect data relevant to the estimand of interest results in a more serious missing data problem with respect to estimating the value of the estimand.

3.2 | Missing data assumptions

In clinical trials, the presence of missing data creates complexity since any analysis requires us to make an assumption about the distribution of the unobserved data. Critically, this assumption is untestable. There are three broad classes of missing data assumptions originally introduced by Rubin in 1976:[26] Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). A missingness mechanism is said to be MCAR where the probability of a datum being missing does not depend on the unobserved value of the datum, or the observed values of other recorded variables. For example, (C)DLQI data in the ADAPT trial would be MCAR if everyone in the trial had an equal probability of having their 24-week outcome recorded. Under MCAR, the marginal distribution of the unobserved data, which expresses the probability distribution of the unobserved data without reference to the values of any other variables, will be the same as the marginal distribution of the observed data. Broadly, missingness is unrelated to the inference we wish to draw.

In the more general case of MAR, the probability of a datum being missing does not depend on the unobserved value of the datum, given observed information. The missingness depends on observed data values marginally, but given the observed data is conditionally independent of the missing data values. In the ADAPT trial, if younger patients with less severe eczema at baseline were more likely to have their (C)DLQI recorded at 24 weeks, but everyone of a given age and baseline severity were equally likely to have their outcome recorded, the (C)DLQI data would be MAR. Under MAR, the marginal distribution of the unobserved data will, therefore, not be the same as the marginal distribution of the observed data, but the conditional distribution of the unobserved data given the observed data will be the same, regardless of whether the data was observed or not. Alternatively, the missingness process is termed MNAR where the probability of a datum being missing does depend on the unobserved value of the datum, even given the observed data. For example, in ADAPT, (C)DLQI would be MNAR if participants with worse (C)DLQI outcome at 24 weeks were more likely to have their 24-week (C)DLQI missing. Here, the marginal and conditional distributions of unobserved values will differ to that of the observed data.

Although MCAR can be distinguished from MAR by a comparison of covariate distributions for observed versus missing outcome values, for example, via a logistic regression, the data at hand cannot confirm which mechanism is operating. Since we can never know what the missing values would have been, we cannot distinguish between MNAR and both MAR and MCAR. Sensitivity analysis thus plays a critical role to reveal the extent to which the results depend on the assumptions. In collaboration with the trial team (including those on-the-ground collecting data and clinical investigators) and/or regulators, we must pick the most plausible assumption for the missing data at hand, which targets the estimand of interest, and conduct primary analysis under that assumption. Sensitivity analyses under alternative plausible missing data assumptions, which also target the same estimand, should subsequently be undertaken to assess the sensitivity of inferences to the underlying assumptions, including those made for missing data. Ideally, inferences would not change across sensitivity analysis, providing reassurance that the missing data did not seriously affect the interpretation of results.[27] If this is not the case, such analysis allows individuals to assess under what conditions results change, and how plausible these conditions are.

3.3 | Approach to primary analysis

The primary analysis of a clinical trial is typically conducted under the assumption of MAR; analysis under MNAR almost always requires external information to be combined with the observed data, which is undesirable for a primary analysis. MAR is a natural starting assumption for the unobserved data because it essentially states that the distribution of a patient's data at the end of the study, given their earlier observed data, does not depend on whether the data at the end of the study were observed. A version of this assumption provides a rationale for inferring that the results of a clinical trial will apply to individuals with similar characteristics from the population of interest not included in the trial. The strong MCAR assumption is often unlikely to be valid in the clinical trial context where drop out may be effected by treatment and observed responses. This is particularly likely in longitudinal settings when data are missing due to uncontrollable events such as receipt of rescue medication, since these events are often associated with the study variables. For example, in ADAPT, it is unrealistic that (C)DLQI is missing just by chance at 24 weeks. It is more plausible that, conditional on baseline, treatment and earlier responses (which may have led to treatment withdrawal or initiation of rescue medication hence the missingness) data are MAR. Although not verifiable, MAR does not require the modeling of the dropout procedure. Under MAR, valid inference can be obtained from the likelihood of the observed data only.[28] MAR will be the primary assumption made for the unobserved data in our analysis of the ADAPT and acupuncture trial.

In practice, when data are missing, we have two accessible alternatives that provide valid inference under MAR:[28,29]

1. Perform a longitudinal likelihood based data analysis, which makes use of all the observed pre-deviation data from each patient, for example, a mixed model for repeated measures (MMRM);
2. Use MI and impute missing data under the primary MAR analysis assumption, fit the primary analysis model (the model of interest which would have been used in the absence of any missing data) on each imputed dataset and use Rubin's rules to combine results for inference.

The two approaches will be approximately equivalent, provided the variables used in the imputation model are the same as those included in the analysis model and conditionals are accommodated by a single joint model.[30] In such settings, MI essentially provides an approximation to the observed likelihood analysis. If an infinite number of imputations could be performed, then the two approaches would be equivalent. In practice, the level of equivalence will depend on the number of imputations due to the Monte Carlo (simulation) sampling variability of the imputation process (described in more detail below), thus will be stronger for a larger number of imputations.[31]

The MI procedure can, however, be a simpler, more practical option when one wishes to include additional "*auxiliary*" information, which is predictive of missingness, within the analysis. Auxiliary variables can readily be incorporated within the imputation model but need not be conditioned on in the analysis model. This is useful when one does not want to estimate the treatment effect conditional on the values of said auxiliary variables, but requires the auxiliary information to justify or strengthen the MAR assumption. For example, in ADAPT, we can incorporate interim follow-up measurements recorded over weeks 4 to 20 in the imputation model and not in the analysis model. If additional data on intercurrent illness post-randomization were recorded and this was thought to be predictive of missingness, this could also be included in the imputation model and not in the analysis model. Option 1 requires careful model specification to ensure the appropriate variables are included in the analysis but not conditioned on. We now expand upon the core principles of the standard MI procedure to provide the necessary background to the sensitivity analysis context and to demonstrate primary analysis under MAR using MI for the ADAPT trial.

3.4 | Multiple imputation

MI was originally introduced by Rubin in 1978.[29,32] MI uses frequentist inference, based on large sample Bayesian arguments for justification. The method and its applications to clinical trials have since been studied extensively by many.[5,30,33-35] The standard MI procedure is conducted under the assumption of MAR and can be broken down into three core stages summarized in Figure 3. In stage 1, missing data are imputed following the Bayesian paradigm by drawing from the posterior predictive distribution of the observed data under the assumption of ignorability (ie, MAR). This is done independently $k = 1, \dots, K$ times to create K completed datasets. Stage 2 proceeds by analyzing each imputed dataset using the substantive analysis model of interest, which would have been used in the absence of missing data. This results

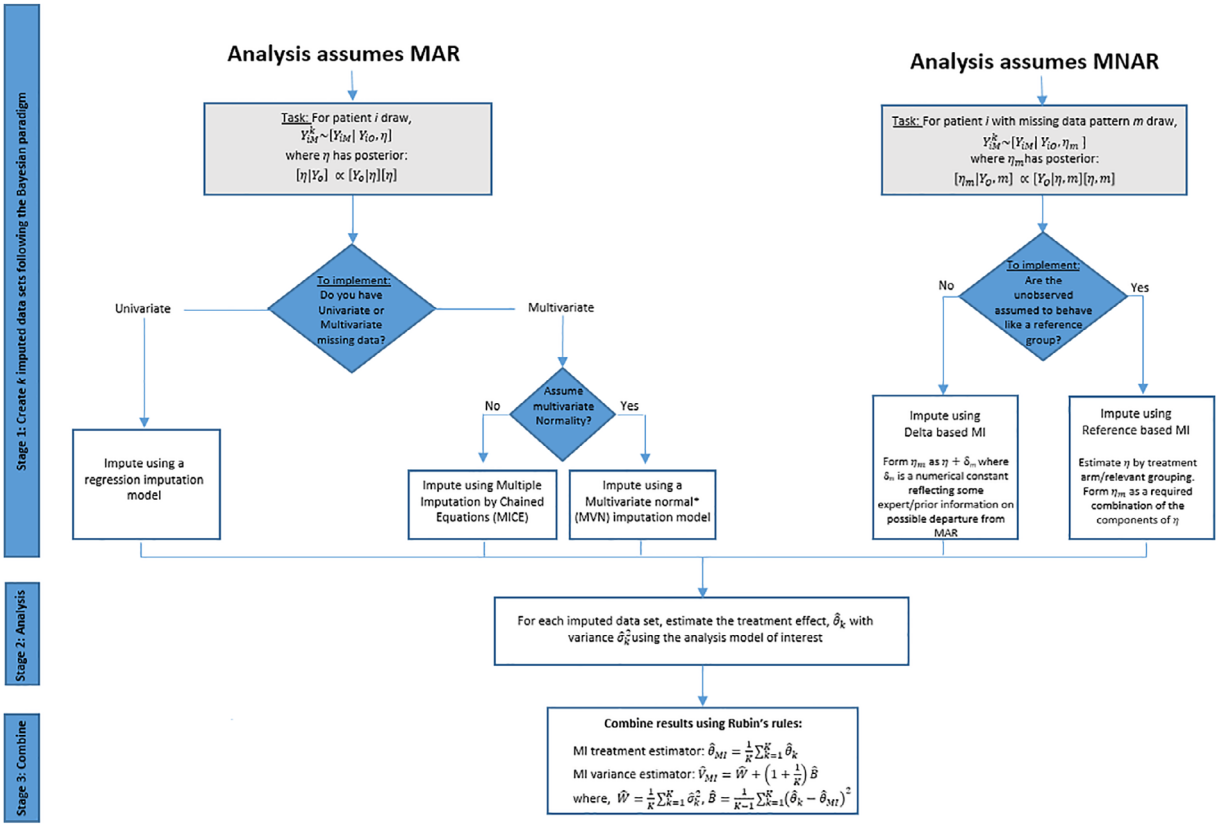

FIGURE 3 Summary of the three core stages of the generic MI procedure for a scalar treatment effect and how to implement the imputation task; within stage 1 we distinguish the conceptual task followed by how this is operationalized [Colour figure can be viewed at wileyonlinelibrary.com]

in K estimated treatment effects, each accompanied by an estimate of variance. In Stage 3, results across imputed datasets are combined using Rubin's rules[29] to give a single MI estimate for inference. The MI treatment estimator, computed using Rubin's rules, is the average of the treatment estimates across imputed datasets. The MI variance estimator consists of the average estimated variance of the treatment effect over the K imputed datasets plus the variance of the treatment effects, $\hat{\theta}_k$, across the K imputations (multiplied by $(1 + 1/K)$ to adjust for finite imputations[29]).

It is important that the imputation model used in stage 1 includes all the variables in the analysis model used in stage 2. This is because imputation-analysis model compatibility is required for unbiased estimation within the analysis stage.[30,36] For clinical trials, the imputation model should, therefore, include the outcome and the treatment allocation in addition to any covariates and interactions that are included in the analysis model. The treatment arm can be incorporated either by conducting imputation separately for each arm (which implicitly fully allows for interactions between all included covariates and randomized group) or by including randomized arm as a covariate in a single imputation model (slightly stronger assumptions since covariate-treatment interactions are fixed to zero). Auxiliary variables that are thought to be predictive of missingness but are not required in the analysis model can also be included in the imputation model.[30,36] Within the analysis of the ADAPT trial, we will incorporate the interim (C)DLQI follow-up measures in the imputation model to render the MAR assumption more plausible.

When conducting MI, a choice must be made about the number of imputations to perform. This will depend on the precision required in estimation. A simple rule of thumb is to use one imputation per percent of missing data.[37] For more critical inferences, more imputations may be required to ensure efficient point estimates and standard errors that would not change if the imputation were to be repeated again.[5] While few imputations can be sufficient to justify the long-run properties like test size, more imputations can increase the power non-trivially. Post imputation, we recommend examination of the Monte Carlo error, which quantifies the sampling variability across imputations, to assess that the number of imputations performed provides an adequate level of precision. White et al[37] recommend (i) the Monte Carlo error of a coefficient should be less than or equal to 10% of its SE, (ii) the Monte Carlo error of the test statistic, coefficient/std.

error coefficient ≈ 0.1, and (iii) the Monte Carlo error of the *P*-value be ≈ 0.01 when the *P*-value is 0.05 and 0.02 when the *P*-value is 0.1. More recently, von Hippel[38] proposed a two-stage procedure to establish the number of imputations required for adequate precision that is more accurate with larger amounts of missing data (>50%). This involves imputing the data with a small number of imputations, then utilizing a quadratic rule[38] to get the required number of imputations for the desired level of precision.

So how do we draw imputations from the Bayesian paradigm in stage 1? As we summarize in Figure 3, under MAR, this will depend on whether there is univariate or multivariate missing data and the types of variables to be imputed. With missingness on a single outcome variable, under MAR, imputations from the Bayesian paradigm can be obtained using a regression model and uninformative priors. For example, suppose for now we ignore the interim follow-up periods in ADAPT and the trial consisted of just baseline and a single follow-up (C)DLQI measure at week 24, which we denote by Y_{i1} and Y_{i2} for patient i. Let $\mathbf{X}_i$ denote the covariate vector containing patient i's treatment allocation and randomization stratification factors (age and IgE) and suppose Y_{i2} is MAR dependent on Y_{i1} and $\mathbf{X}_i$. We can construct an imputation model as a regression model of the week 24 outcome (Y_{i2}) on baseline outcome (Y_{i1}), treatment, and randomization stratification factors ($\mathbf{X}_i$) fitted to the observed data. Imputed datasets are created by repeatedly drawing the regression parameters from their posterior distribution (using an uninformative prior), followed by the missing data from the posterior predictive distribution using the current parameter draw (unique to each imputation step) as follows,

1. Regress Y_{i2} on Y_{i1} and $\mathbf{X}_i$ using the complete records: $Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \boldsymbol{\beta}_2 \mathbf{X}_i + e_i, e_i \sim N(0, \sigma_{2.1})$, to obtain $\hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\beta}}_2$ and $\hat{\sigma}_{2.1}$,
2. For imputation k, using uninformative priors, draw $\hat{\beta}_0^k, \hat{\beta}_1^k, \hat{\boldsymbol{\beta}}_2^k$ and $\hat{\sigma}_{2.1}^k$ from the Bayesian posterior of $\hat{\beta}_0, \hat{\beta}_1, \hat{\boldsymbol{\beta}}_2$ and $\hat{\sigma}_{2.1}$,
3. Draw missing data from: $Y_{i2,k} = \hat{\beta}_0^k + \hat{\beta}_1^k Y_{i1} + \hat{\boldsymbol{\beta}}_2^k \mathbf{X}_i + e_i, e_i \sim N(0, \hat{\sigma}_{2.1}^k)$,
4. Repeat steps 2 and 3 K times.

With missingness on a single noncontinuous variable, a logistic, multinomial, or ordinal regression model may be used. With multivariate data subject to nonresponse and a monotone missingness pattern, that is, when the measurements can be ordered such that missingness on one variable implies missingness on others, imputation can also proceed using regression modeling. We now do not ignore the interim follow-up time points in the ADAPT study. Figure 1 illustrates how the observed missingness pattern is monotone in the ADAPT trial—where patients miss a (C)DLQI measurement, they are missing (C)DLQI at all the subsequent time points. For monotone missing data, we can impute variables in increasing order of missingness using, for each time point, a regression model that includes the completed variables only. For imputation at each subsequent time point, previously imputed variables are included in the imputation model. For example, in ADAPT, where interest lies in the imputation of the (C)DLQI outcome, we can impute missing (C)DLQI values as follows.

Variable to impute:	Covariates included in regression model:
Week-4 (C)DLQI	Age, treatment, baseline (C)DLQI
Week-8 (C)DLQI	Age, treatment, baseline (C)DLQI, week-4 (C)DLQI
Week-12 (C)DLQI	Age, treatment, baseline (C)DLQI, week-4 (C)DLQI, week-8 (C)DLQI

This process can be readily implemented in Stata using the `mi impute monotone` command. The analysis model of interest in ADAPT is a regression of the week 24 (C)DLQI measure on treatment group, baseline (C)DLQI, and the randomization stratification factors of age and IgE. The variables in the imputation model will, therefore, be the same as those in the analysis model along with the interim (C)DLQI measures taken over week 4 to week 20 to strengthen the MAR assumption. A total of 50 imputations will be performed, which we will confirm provides adequate precision.

The ADAPT dataset is in wide format with one observation per individual, where `id` is the unique individual identifier and `treat` is the randomized treatment assignment, to placebo (`treat` = 0) or active (`treat` = 1). `CDLQI_wj` for $j = 4, 8, 12, 16, 20, 24$ is the post-baseline (C)DLQI measurement at time j (weeks). `CDLQI_1` is the baseline (C)DLQI measurement and `agestrat` is baseline age and `igestrat` baseline IgE. To conduct MI within Stata, we first declare the desired style of MI data to be produced (style `flong` produces imputed datasets stacked sequentially below the original data) and register the variables to be imputed. We then perform MI under MAR using monotone regression imputation using the `mi impute monotone` command as follows,

```
· mi set flong
· mi register imputed CDLQI_w4 CDLQI_w8 CDLQI_w12 CDLQI_w16 CDLQI_w20 CDLQI_w24
· mi impute monotone (regress) CDLQI_w4 CDLQI_w8 CDLQI_w12 CDLQI_w16 CDLQI_w20
  CDLQI_w24 = i.agestrat CDLQI_1 i.treat i.IgEstrat, add(50) rseed(2301)
```

Analysis of the imputed datasets and the combination of results using Rubin's rules (including Monte Carlo error) is then conducted using the `mi estimate` command as follows,

```
· mi estimate, mcerror: regress CDLQI_w24 i.treat i.agestrat CDLQI_1 i.IgEstrat
```

The obtained MI (MAR) treatment effect estimate is −3.76, SE = 1.57, 95% CI −6.94, −0.58, $p = 0.022$. The inclusion of the `mcerror` option in the estimation step outputs the Monte Carlo error of the treatment estimate to confirm that we have an adequate number of imputations. The Monte Carlo error in the treatment effect is 0.11 (approx 3%); the Monte Carlo error of the test statistic is 0.08 and of the *P*-value is 0.004, which implies that 50 imputations was adequate: We have an appropriate level of precision. If greater precision was required, a greater number of imputations could be produced.

Generally, under a non-monotone missingness pattern, forming an imputation model for multivariate data is more complex. In such settings, two main routes have been established for imputation. The first, "Multivariate Imputation by Chained Equations" (MICE), or "Fully Conditional Specification" (FCS), involves a series of univariate conditional models, formed as a regression of each partially observed variable, given all other variables.[39] Each univariate imputation model is specified according to the characteristics of the variable being imputed. For example, binary variables can be modeled using a logistic regression and continuous variables modeled using linear regression. A short-circuited Gibbs-type sampling procedure is used to impute variables. We refer the reader to Van Buuren[36,39] and White[37] for more technical details on the MICE or FCS approach to MI. Within Stata, the `mi impute chained` command can be used to conduct MICE.

As an alternative to MICE, a joint multivariate normal (MVN) model can be assumed for all variables. Missing values are then imputed using an iterative Markov Chain Monte Carlo (MCMC) method. This involves forming an initial estimate of the multidimensional parameter of this MVN distribution; missing data are drawn from the appropriate conditional distribution using the previously estimated parameter; a new value of the multidimensional parameter of the joint MVN data distribution is then drawn from its complete-data posterior given the newly imputed data. The process is repeated until appropriate convergence is satisfied. The number of iterations required to reach convergence is often referred to as the *burn in*. Upon convergence, the current draw of missing data is retained to form the first imputed dataset (along with the observed data). So that subsequent imputed datasets are independent, once convergence has been obtained, and the first draw of missing data is retained, a number of iterations of the MCMC process are taken prior to retaining the next draw of missing data to form the second imputed dataset. This number is often known as the *burn between*. The updating of the chain, the burn-between, and collection of the imputed data, is repeated K times. We refer the reader to Schafer[35] (pp. 306-309) for more technical details on the MCMC procedure. In practice, we can conduct imputation under MAR using the MVN distribution using inbuilt MI commands in statical software. In Stata, the `mi impute mvn` command can be used to perform MI via this route.

Using a joint MVN model for MI does necessitate strong assumptions of normality. Schafer, however, reports simulations that show imputations drawn under the MVN model are robust to moderate skewness.[35] Additionally, Schafer reported the normal model to be a useful tool for imputing ordinal and binary data when no category has prevalence below 10%.[35] Nominal variables can be included in the model with a series of binary dummy variables. Thus, in practice, the normal model is useful even when the data are not normal. As previously discussed, assumptions of normality can be avoided with MICE, thus this option can be used if the MVN approach is not considered appropriate. In practice, for analysis under MAR, only MI via monotone regression modeling, MICE, or MVN is required. In Section (3.5), we will see that MI can also be used to explore departures from MAR, that is, for analysis under MNAR, avoiding the need to fit such MNAR models directly.

3.5 | Sensitivity analysis

Procedures for sensitivity analysis require the measurement process and missingness mechanism to be jointly modeled. For example, for ADAPT, we need to model not only the (C)DLQI outcome, but also incorporate a model for the

mechanism causing missingness in (C)DLQI within sensitivity analysis. There are two ways this can be done. First, we can specify a model for the missing status for (C)DLQI, given the measurement data, with a marginal model for the data. For example, a logistic regression model could be used to model the probability of (C)DLQI being missing at week 24, with a parameter that governs how this depends on the unobserved outcome, fitted alongside a model for the (C)DLQI data. Alternatively we can specify the conditional distributions of the (C)DLQI response data given the fully observed data with a marginal model for the missingness process. For example, we can specify a MVN model for the unobserved data, which has mean higher by a certain proportion than the observed data. The former factorization is referred to as the *selection model*, while the latter is the *pattern-mixture model*.

For sensitivity analysis to be useful, methods and assumptions must be transparent and interpretable to all involved in the trial, not just the experienced statistician. If the assumptions cannot be understood, then neither can the results. We believe that assumptions for the unobserved data are more accessible when expressed in the pattern-mixture form. This approach explicitly specifies how the unobserved data differs from the observed.

To see this formally, in the following, we group planned outcome measurements for patient i at times $j = 1, .., J$ into the vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iJ})$, and denote the vector of observed covariates, including treatment assignment, as $\mathbf{X}_i$. The distribution of the measurement data, for patient i, is defined as $f(\mathbf{Y}_i|\mathbf{X}_i, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the key vector-parameter of interest. We define response indicators for each planned measurement, R_{ij}, as $R_{ij} = 1$ if Y_{ij} is observed or $R_{ij} = 0$ if Y_{ij} is unobserved. For each patient, the response indicators can be represented by the vector $\mathbf{R}_i = (R_{i1}, \dots, R_{iJ})$. The missing data mechanism is defined as the vector process generating $\mathbf{R}_i$ and is modeled as the conditional distribution, $f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is the vector-parameter for the missing data mechanism, which quantifies how the unobserved data depends on the observed and unobserved data values. $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are assumed to be separate/distinct. As shown by Little and Rubin,[28] the two different ways the joint distribution of the data and missingness mechanism can be factorized are,

$$f(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i, \boldsymbol{\phi}) f(\mathbf{Y}_i|\mathbf{X}_i, \theta) = f(\mathbf{Y}_i, \mathbf{R}_i|\mathbf{X}_i, \theta, \boldsymbol{\phi}) = f(\mathbf{Y}_i|\mathbf{R}_i, \mathbf{X}_i, \theta) f(\mathbf{R}_i|\mathbf{X}_i, \boldsymbol{\phi}). \quad (1)$$

Either, as expressed on the left, a model for the missing data mechanism, $\mathbf{R}_i$ given the measurement data, with a marginal model for the data $\mathbf{Y}_i$ can be specified—the selection model—or the conditional distributions of the response data given the fully observed data on the right with a marginal model for the missingness process can be given—the pattern-mixture model. The pattern mixture model requires specification of the joint distribution of the partially and fully observed response variables $\mathbf{Y}_i$, for each pattern of missing data, which implies the conditional distribution of unobserved response data ($\mathbf{Y}_{iM}$) given the observed response data ($\mathbf{Y}_{iO}$) within each pattern as follows,

$$f(\mathbf{Y}_i|\mathbf{R}_i, \mathbf{X}_i, \theta) f(\mathbf{R}_i|\mathbf{X}_i, \boldsymbol{\phi}) = f(\mathbf{Y}_{iM}|\mathbf{Y}_{iO}, \mathbf{R}_i, \mathbf{X}_i, \theta) f(\mathbf{Y}_{iO}|\mathbf{R}_i, \mathbf{X}_i, \theta) f(\mathbf{R}_i|\mathbf{X}_i, \boldsymbol{\phi}).$$

In the pattern mixture framework, the assumptions correspond directly to what is observed, that is, the unobserved data of the deviating patients in the ADAPT trial having different distributions. Pattern mixture models are "underidentified" by construction,[40] meaning the observed data do not reveal the distribution of the unobserved data. Hence, to use such models in practice, additional assumptions are introduced that allow unidentifiable parts of models to be identified from other groups of subjects, that is, in ADAPT, we need to specify exactly how the unobserved (C)DLQI data for the deviators differs to those observed.

In the selection modeling framework, the missing process is alternatively explicitly modeled alongside the observed data. Selections models are intuitively nice since the factorization matches how we imagine data as being generated: full data exist then missing data happen. But, in practice, when we have data, it is often easier to think about the distribution of the unobserved data and how this differs to that observed; the pattern-mixture modeling approach. In the selection modeling framework, assumptions are typically framed around the anticipated odds of response, given baseline covariates and the response measurement. We have found that expressing differences in the odds of response per unit change in the response, conditional on other variables in the analysis model, is less understandable to clinical colleagues. However, as expanded upon further below in Section 4.3, it is not necessarily always easier to interpret assumptions in the pattern mixture framework. But, it is the modeling approach adopted here since it readily allows for accessible analysis using MI.

In individual trials, there will be numerous ways in which a pattern mixture model (or selection model) can be fully specified; however, many specifications will be practically implausible. A commonly advocated principled way to perform sensitivity analysis in the clinical trial setting is to explore departures from the joint distribution implied by MAR.[41,42] MAR provides an unambiguous starting point for MNAR exploration.

As discussed by Daniels and Hogan,[42] starting with specification of the conditional data distribution implied by MAR, one can readily perform sensitivity analysis exploring departures from MAR by specifying a higher or lower mean outcome value for the individuals with unobserved data. After specifying and fitting the separate response models for each pattern, the models are weighted by their respective probabilities to obtain inference. Either a maximum likelihood (ML) or full Bayesian approach can be used to fit pattern mixture models. A detailed example showing how to implement a full Bayesian approach to pattern mixture modeling using `winBUGS` is presented by White et al.[43] Alternatively, MI provides an accessible solution.

The pattern-mixture model readily allows for missing outcomes to be imputed under a chosen scenario. As reviewed in Section 3.4, the standard MI procedure imputes missing data using the conditional distributions of partially observed response data given the observed response data under the assumption of MAR. Following a pattern mixture framework, we modify the conditional distributions implied by MAR within each missing data pattern as appropriate. The modified conditional distributions are then used within the MI algorithm, in place of the MAR distribution to impute under MNAR. The imputed data are analyzed using the primary substantive analysis model, which would have been used in the absence of any missing data. When MI is performed in such a manner, this has been termed "*Controlled MI*."[16] MI avoids the need to fit the pattern-mixture models directly, which can often be quite complex and require sophisticated one-off programing. MI can be a more practical and appealing approach for busy trialists.

In the next section, we outline and illustrate how to conduct controlled MI using a pattern-mixture approach where the difference between the MAR and MNAR distribution is described using a specified numerical parameter, termed "δ-based MI."

4 | SENSITIVITY ANALYSIS USING Δ-BASED MI

4.1 | Sensitivity analysis of the ADAPT trial

As in the ADAPT trial, it is frequently of interest to explore the possibility of the unobserved having a poorer response than those observed; in other cases it might be of interest to explore the impact of the unobserved having a better response than those observed. δ-based MI provides a useful accessible route for exploring such departures from MAR in clinical trials. In the most simple case, we specify a single numerical parameter, δ, which governs the mean difference between the MAR and MNAR distribution for all individuals with missing data. For continuous data, we propose the unobserved data has a distribution with mean either higher or lower than that implied by the observed data and MAR, and impute under this distribution. For linear models, δ-based MI can be implemented by imputing under MAR, then adding/subtracting the constant δ directly onto the resulting imputed values to increase/lower the mean response below that predicted under MAR. The imputed datasets can then each be analyzed; we have injected the desired MNAR mechanism in the imputation step. In more complex scenarios, where one wishes to vary the unobserved data distribution by missingness pattern, which may depend on time of drop out, reason for missingness or an alternative patient characteristic (eg, treatment arm), $\boldsymbol{\delta} = (\delta_1, \dots, \delta_m)$ may be a vector parameter quantifying the difference in the means of the MAR and MNAR distributions, for missing data patterns $1, \dots, m$.

We assume throughout that δ, or $\boldsymbol{\delta}$, represents an offset term, or collection of offset terms, which describes the difference in the mean outcome between the observed and unobserved cases. The offset term does not necessarily have to be a mean shift in the intercept and additional offset terms, which represent a shift in slope for specified covariates can in theory be introduced. However, this creates additional complexities, which we do not discuss further here.

We now demonstrate δ-based MI for the ADAPT trial and explore the impact of a different mean response among the unobserved at the week 24 follow-up visit. We first consider the most simple scenario where we specify a single numerical parameter, δ, to govern the difference between the MAR and MNAR distribution for imputation. It is likely that participants week 24 responses had a worse (higher) (C)DLQI score than for those observed with similar characteristics in the absence of rescue mediation. Furthermore, for those who withdrew from treatment, since they decided not to progress, it is also likely they would have experienced a worse response than for those observed.

We define δ as the fixed difference in (C)DLQI outcome between observed and unobserved cases at week 24. For each patient with missing data, we then modify the MAR imputed observations at week 24 by δ. We will repeat the analysis for a range of δ values corresponding to 25%, 50%, 75%, and 100% of the absolute change of the (C)DLQI observed over 24 weeks in all participants.

TABLE 1 Estimated 24 week treatment effect on C(D)LQI in primary and sensitivity analysis of the ADAPT trial

Analysis	Treatment Est.	95% conf. int.	Std. Err.	*P*-value
Primary (MI MAR)	−3.76	−6.94 to −0.58	1.57	0.022
Sensitivity (MI $\delta = 1.875$)	−4.11	−7.35 to −0.87	1.60	0.014
Sensitivity (MI $\delta = 3.75$)	−4.46	−7.81 to −1.11	1.66	0.010
Sensitivity (MI $\delta = 5.625$)	−4.81	−8.31 to −1.31	1.73	0.008
Sensitivity (MI $\delta = 7.5$)	−5.16	−8.85 to −1.48	1.83	0.007

We have already performed imputation under MAR for ADAPT in Section 3.4 (results in Table 1). To perform the sensitivity analysis, we start by using this MAR imputed dataset. Over the 24-week treatment period, the observed overall unadjusted mean change in (C)DLQI was −7.5. In the first sensitivity analysis, we will increase the MAR imputed (C)DLQI values at week 24 by 1.875 (25% of the unadjusted mean change in (C)DLQI). The code required to complete these steps is as follows.

```
· use adapt_MAR, clear
· mi passive: generate byte imputed=_mi_miss
· replace imputed = 1 if imputed==.
· generate float CDLQI_w24_Delta1 = CDLQI_w24
· replace CDLQI_w24_Delta1=CDLQI_w24 +1.875 if imputed==1
```

The analysis model from the primary analysis is retained and fitted to the updated imputed data using the usual `mi estimate` command in Stata as,

```
· mi estimate, mcerror: regress CDLQI_w24_Delta1 i.treat i.agestrat CDLQI_1
  i.IgEstrat
```

We repeat the above sensitivity analysis using increasingly larger values for delta of 3.75, 5.625, 7.5, corresponding to 50%, 75%, and 100% of the overall observed unadjusted mean change in (C)DLQI. To do so small edits are required to the above two extracts of code. For example, for the second sensitivity analysis, we create a new variables called `CDLQI_w24_Delta2` in place of `CDLQI_w24_Delta1` and use the appropriate updated delta adjustment in place of 1.875. Results of all the sensitivity analyses conducted are shown in Table 1.

When we assume the individuals who received rescue medication or withdrew from treatment had a worse response than that predicted under MAR (if they had remained in the trial on-treatment), we see the treatment effect marginally increases. When we assume the unobserved week 24 response is higher by 7.5 points than that predicted under MAR, the obtained treatment effect is −5.16, Std Err = 1.83, and $p = 0.007$. Thus, assuming a fixed worse difference in response among those with missing data results in a larger treatment effect estimate. Since more patients in the placebo arm deviated this is unsurprising. However, the increase is not dramatic and does not change the overall interpretation. The sensitivity analysis confirms our conclusions from the primary analysis are robust to alternative plausible missing data assumptions.

4.2 | Incorporating the missing data pattern in a longitudinal trial setting

The above δ-based analysis for ADAPT assumed a fixed difference in response between the observed and unobserved cases at the final follow-up. Such an approach can be used in a trial with baseline and a single follow-up only or in longitudinal settings where it is appropriate to assume a fixed difference in outcome regardless of time or reason for drop out. For ADAPT, it was considered most plausible to assume a fixed difference in outcome regardless of the missing data pattern or missingness reason. In other scenarios, it may be more appropriate to vary the parameter governing the difference between the MAR and MNAR distribution, δ, by missing data pattern, which may depend on time of drop out, treatment group and/or reason for missingness.

One can define for each missing data pattern m a different anticipated difference in the mean response from MAR. In such settings, δ-based MI will require the specification of many sensitivity analysis parameters as thus it will be required to define $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ as a collection of sensitivity analysis parameters for each missing data pattern m. For example, in the ADAPT trial, a different δ adjustment could have been made for (i) the individuals who received rescue medication and (ii) the individuals who withdrew from treatment. We provide example code in Appendix A for implementing this. Furthermore, the time of deviation (withdrawal or receipt of rescue therapy) could have been incorporated and different adjustments at week 24 made for those who received rescue medication/withdrew at week 8, 12, 16, 20, and 24. Naturally, if time of deviation is taken into account, this can entail the specification of many parameters which can be practically challenging to form.

A fixed departure from MAR, regardless of time of drop out, may often not be credible. In such cases, it may be alternatively appropriate to explore a change in slope post-deviation and so to define δ to be a scalar sensitivity analysis parameter representing the change in rate of response post-deviation relative to MAR. This will reduce the number of parameters to be specified. In such settings, one can add δ to the first post-deviation imputed value, 2δ to the second, and so on. We note here that if the analysis model only incorporates the last observed time point (like in ADAPT), then we may only need to actually edit that response accordingly by the appropriate multiple of δ. This approach is outlined by Carpenter and Kenward.[5,41] In such settings, one may wish to vary δ (defined as the change in slope post-deviation) by treatment arm, reason for missing data or an alternative grouping. Such an approach that reduces the number of parameters will inevitably be a simplification of reality. Additional example code on how to implement a difference in the rate of change in (C)DLQI for the ADAPT trial, with and without additional variation by deviation reason (withdrawal/rescue therapy), is given in Appendix A.

Since the analysis model for ADAPT only incorporated baseline and the week 24 outcome we did not alter imputed values at earlier follow-up time points. But in different scenarios where the analysis model includes earlier follow-up data, for example, an MMRM, an additional consideration is whether earlier missing values should also be edited to reflect a departure from MAR. If required, a delta adjustment can also be made to earlier missing values. We have also not yet discussed interim missing data, which is when individuals have missing data at some point in the follow-up but data are observed later. Interim missing observations may often be reasonably imputed under MAR. But in some circumstances, dependent on the clinical context, it may be appropriate to also alter interim missing values using a delta adjustment.

4.3 | Specification of δ

Formally using the same notation as introduced in Section 3.5, for MNAR imputation, for each deviating individual i with missing data pattern m_i, we specify the distribution of their missing outcomes, given their observed data as,

$$\mathbf{Y}_{Mi}^{k} \sim \left[\mathbf{Y}_{Mi} | \mathbf{Y}_{Oi}, \boldsymbol{\eta}_m\right]. \tag{2}$$

Here, $\boldsymbol{\eta}_m$ represents the vector of parameters of this distribution, whose values differ across missingness patterns m, and whose values we first have to form for each imputation k, before we can impute missing data from Equation (2) using the standard MI procedure. The parameters of the imputation model under MAR, $\boldsymbol{\eta}$, are obtainable using the observed data. For δ-based approaches generally, for each missing data pattern m, the parameters $\boldsymbol{\eta}_m$ are constructed using a draw of the parameters $\boldsymbol{\eta}$ of the MAR implied conditional distribution and a numerical sensitivity analysis parameter. This numerical sensitivity analysis parameter governs the degree of departure from MAR and creates the shifted distribution for imputation. Formally, the postulated numerical difference in the mean parameter for the observed data and missing data pattern m, that is, the sensitivity analysis parameter is denoted by δ_m. The current draw of η for imputation k is then edited accordingly (by adding/subtracting δ_m) to obtain $\boldsymbol{\eta}_m$ for imputation. The MI algorithm then proceeds using the shifted distributions. We note that $\delta_m = \delta$ in the above example for ADAPT, as will be relevant in many other settings.

In any trial setting, the extent to which the parameters of the distribution of missing data are likely to differ from the observed for each missing data pattern requires carefully consideration. What value or values to use for the sensitivity analysis parameters might not be immediately obvious. The first point to consider is whether it is most likely that deviators had a poorer/better response than those observed? Then, in the direction(s) of interest, one needs to consider by how much does the response likely vary for each missing data pattern. It is of utmost importance to carefully consider what values of δ (or δ_m for missing data pattern m) represent a plausible degree of departure from MAR in the specific setting

at hand. A discussion between clinicians, researchers on the ground interacting with the participants, and statisticians should occur to ensure appropriate values are chosen. The scale of the outcome will of course feed into such discussions around the degree of departure from MAR. Data from similar studies may also provide useful insights into realistic values alongside experts' opinions.

A pragmatic approach is to use the observed data to inform realistic values of the sensitivity analysis parameters. For example, in ADAPT, it was pre-specified for the primary outcome (objective SCORAD), which was collected longitudinally, that sensitivity analysis would explore the impact of a range of fixed mean differences in outcome for the unobserved regardless of deviation pattern (δ) corresponding to 10%-50%, the observed rate of change over 24 week in all participants.[22] One might alternatively consider using the minimum clinically important different for the outcome under study (often also used to derive the trials sample size) to guide a worse/best case scenario. A δ-based adjustment employing the minimum clinically important different or a multiple of this might reveal whether in a more extreme setting results vary.

An alternative approach to δ-adjusted sensitivity analysis progressively increases δ-based adjustments from 0 until the conclusions from the primary analysis are overturned. Each δ represents an increased departure from MAR ($\delta = 0$). If the value of δ that changes the conclusions is implausible, that is, realistically the missing data would not be that different from the observed data, then greater confidence in the primary results can be inferred. This is referred to as a "tipping point analysis" by Yan.[44] Ideally a range of acceptable values of δ should be agreed within trial teams a-priori when taking this approach. We caution against such tipping point approaches if careful thought has not been given to plausible values of delta a-priori since the results of the analysis might knowingly or unknowingly influence the subsequent interpretation of the sensitivity analysis.

To summarize, in any trial setting, careful thought must be given to what are appropriate values for the sensitivity analysis parameters when utilizing δ-based MI. For longitudinal trials with a number of different missing data patterns, as with all pattern-mixture approaches and indeed selection modeling approaches, this can require the specification of many parameters. Next we will see that explicit parameter specification is, however, not always required when using controlled MI.

5 | SENSITIVITY ANALYSIS USING REFERENCE-BASED MI

5.1 | Sensitivity analysis of the headache trial

Statements about unobserved patient data can alternatively be made by reference to other groups of individuals in the trial (typically individuals in different treatment arms). That is, the difference between the MAR and MNAR distribution can be described entirely using within trial information by reference to other groups in the data. The parameters of the observed data distribution, estimated assuming MAR, can be mixed around, *across* arms rather than within, to form contextually relevant MNAR distributions for the unobserved data. Data can then be imputed from the conditional distributions pieced together from the MAR parameters. This is referred to as "reference based MI." Reference based approaches, which follow the rule of using within trial information, avoid the need for explicit parameter specification, which can be difficult. In-study data is used to make qualitative rather than quantitative missing data assumptions based on plausible clinical scenarios. The imputed datasets are then analyzed as usual with MI and results combined using Rubin's rules. The primary analysis model, based on a comparison of the randomized groups, is retained in the sensitivity analysis in keeping with the Intention-to-treat (ITT) principle. This allows for the exclusive assessment of the impact of alternative sampling behavior on the primary analysis as originally planned.

There are many ways in which distributions for the unobserved data can be constructed using internal trial parameters. The appropriate choice will be context specific. Some possibilities (not-exhaustive) are:

- In a two arm trial of an active versus a reference treatment, we could impute missing data assuming patients jump to behave like those in a specified reference arm following their last observed time point. This may be suitable when any treatment effect would be expected to stop following drop out. This has been termed *jump to reference* imputation. For example, in the acupuncture versus standard care headache trial, we could impute missing data assuming patients jump to follow the behavior of the standard care arm following their last observed time point. This scenario, considered the most plausible sensitivity analysis for the headache trial, is illustrated schematically in the top right panel of Figure 4.

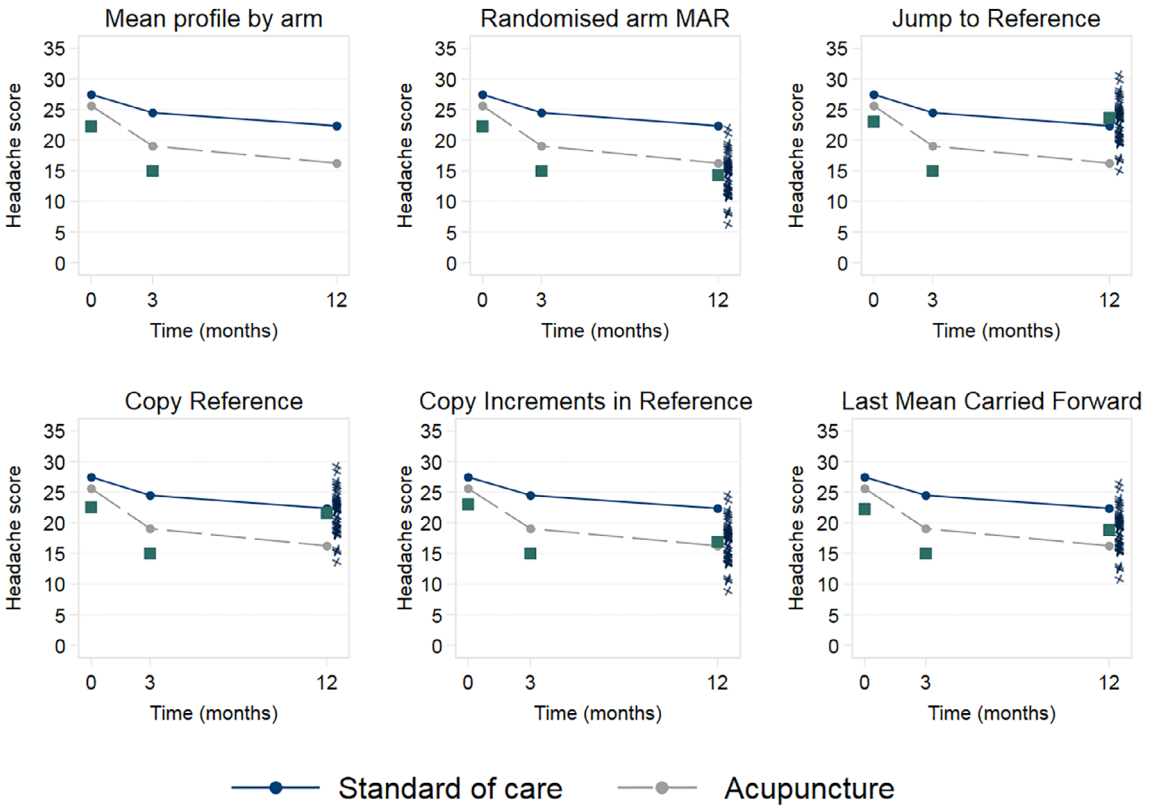

FIGURE 4 Example reference based imputation models for the acupuncture trial. The squares at time 0 and 3 are observed values for a participant in the acupuncture arm who withdrew after the 3 month visit. The black arrows represent the imputation distributions. The squares at time 12 are the mean of the imputed values for that participant in the given reference based scenario. The crosses at time 12 represent the individual imputed values around that mean across 50 multiply imputed datasets for the withdrawing active participant. The reference arm is the standard care arm. This is not an exhaustive display of the MNAR options possible within the reference-based framework [Colour figure can be viewed at wileyonlinelibrary.com]

- Alternatively, we could impute assuming patient outcomes follow the mean increments observed in a reference arm following their last observed time point. This is referred to as *copy increments in reference* (CIR) imputation. Within the acupuncture arm, we could impute assuming the differences in patients mean outcomes over time follow those observed in the standard care arm (bottom row center panel of Figure 4).
- A third option is to impute assuming patients behaved as if they were in a specified reference arm for the full duration of the trial, known as *copy reference* (CR) imputation. In the acupuncture trial, we could impute assuming patients followed the standard care arm behavior for the full trial duration (bottom row left panel of Figure 4), a natural option when we believe patients followed a different (reference) treatment from their randomized allocation throughout the trial.
- Alternatively, *last mean carried forward* (LMCF) imputes assuming patient behavior stays at the mean level for their randomized arm at their last observed time point, appropriate when we believe the effect of randomized treatment is maintained on average over time. This is a more principled version of the classic LOCF analysis (bottom right row panel, Figure 4).

Table 2 summarizes these five imputation options proposed by Carpenter et al[11] for a continuous outcome. The options in Table 2 are not an exhaustive listing of reference based options. Naturally under any reference based method discussed above, for patients in the designated reference arm, their data will be imputed as under randomized arm MAR. Using data from the acupuncture headache trial, we will demonstrate how reference based MI can be accessibly conducted in Stata using the `mimix` command under these five assumptions.[45] Further technical details on the underlying reference based algorithm are provided in Appendix B.

We will first conduct primary analysis under the most plausible randomized-arm MAR assumption for the unobserved data using the reference based MI algorithm of Carpenter et al.[11] The `mimix` command can be downloaded within Stata

TABLE 2 Examples of reference based multiple imputation options

Method	Description
Randomized-arm MAR	Impute assuming patients follow the behavior of their randomized arm. The joint distribution of patients' pre- and post-deviation outcome data is MVN with mean and covariance matrix from their randomized arm.
Jump to reference (J2R)	Impute assuming patient behavior jumps to that of a specified reference arm. The joint distribution is MVN with mean vector from the patients randomized arm up to their last observation time, post-deviation the mean vector follows that observed for a reference group (typically control). The covariance matches the randomized arm for pre-deviation measurements and the reference arm for the conditional components of post- given pre-deviation measurements.
Last mean carried forward (LMCF)	Impute assuming patient behavior remains at the mean level for their randomized arm at their last observed time point. The joint distribution is MVN with mean vector from the patients randomized arm up to their last observation time, post-deviation the means are set equal to the marginal mean for the patients randomized arm at their last observed time. The covariance matrix remains as that for their randomized treatment arm.
Copy increments in reference (CIR)	Impute assuming patient behavior follows the mean increments observed in a specified reference arm. The joint distribution is MVN with mean vector from the patients randomized arm up to their last observed time, post-deviation the patients mean increments follow those from a reference arm. The covariance is the same as in J2R. Appropriate when we wish to assume that post-deviation the disease resumes the course observed in the reference arm.
Copy reference (CR)	Impute assuming patients follow the behavior of a specified reference arm for the duration of the trial. The joint distribution of patients' pre- and post-deviation outcome data is MVN with mean and covariance matrix from a reference arm regardless of deviation time.

by typing `ssc install mimix`. A detailed specification of the commands options is available in Cro et al.[45] A freely available SAS macro by Roger[46] called `miwithd` also implements the algorithm of Carpenter et al. The "five-macros" SAS package, which is a more developed version of the `miwithd` macro, is also available for analysis within SAS.[47] Both SAS implementations are available for download at www.missingdata.org.uk.

Within the headache dataset, `id` is the unique individual identifier and `treat` is the randomized treatment assignment to standard care (`treat = 0`) or acupuncture (`treat = 1`). Baseline covariates include the randomization stratification factors of `age`, `sex`, `migrane` (diagnosis of migraine or tension-type), and `chronicity` (number of years of headache disorder). `head_base` is the baseline headache score. `head` is the post-baseline headache score and `time` is the time of the headache measurement in months (3 or 12 months). The dataset is in "long" format, with one observation per individual per time point, as `mimix` requires.

We impute under MAR and create 50 imputations using an MCMC burn-in of 1000 and burn-between of 500 iterations as recommended by Carpenter and Kenward (p. 84).[5] We include the randomization stratification factors of `age`, `sex`, `migrane`, and `chronicity` in the imputation model. The baseline headache measure (`head_base`) is also included in the imputation model as a covariate, but if this fully observed variable were used as an outcome in the imputation model, the imputation results would be stochastically identical. We include `head_base` as a covariate here so that it will be treated as a covariate in the analysis step; we use the `regress` option to specify that the substantive analysis is a linear regression of 12-month headache score (final time point) on randomized treatment and the included covaraites (`age`, `sex`, `migrane`, `chronicity`, and `head_base`). The `regress` option fits a linear regression of the outcome at the final time point on treatment arm and any covariates included in the imputation model post MI to each imputed dataset, then combines results using Rubins' rules. If an alternative substantive model of interest were required, the `regress` option is not required and the imputed datasets can instead be saved for use with a different analysis model. Although possible, we caution against the use of an analysis model that has variables or structure (eg, interaction terms) not included in the imputation process because this will create additional imputation-analysis model incompatibility. As discussed in Section 3.4, when performing MI under MAR, it is important that the imputation model includes (at a minimum) all

TABLE 3 Sensitivity analysis results for the headache trial using reference based MI with $K = 50$ imputations

Analysis	Treatment Est. (L)	95% CI	SE	*P*-value
Primary analysis				
Randomized-arm MAR	−4.97	−7.40 to −2.54	1.23	< 0.001
Sensitivity analysis[a]				
Jump to standard care	−3.32	−5.70 to −0.94	1.21	0.006
Copy increments in standard care	−3.74	−6.07 to −1.41	1.18	0.002
Copy standard care	−3.80	−6.12 to −1.49	1.18	0.001
Jump to acupuncture	−3.00	−5.44 to −0.56	1.24	0.016
Copy increments in acupuncture	−3.50	−5.91 to −1.10	1.22	0.004
Copy acupuncture	−3.48	−5.87 to −1.09	1.21	0.005
Last mean carried forward	−4.94	−7.38 to −2.50	1.24	< 0.001

[a]Sensitivity analysis results are sorted in terms of what we considered the most plausible assumption (higher position vertically) to least plausible assumption (lower position vertically).

the variables to be included in the analysis model to ensure unbiased estimation. Although the imputation and analysis model will not in fact be fully compatible in reference based settings, the implications of which we expand further on in Section 6, the imputation model should include all variables to be included in the analysis. The `saving` option specifies that the imputed datasets will be saved in a Stata data file called `head_mar`. The commands required are as follows.

```
· use acupuncture, clear
· mimix head treat, id(id) time(time) covariates(age sex migraine chronicity
  head_base) method(mar) m(50) regress clear seed(23) burnin(1000) burnbetween(500)
  saving(head_mar)
```

For sensitivity analysis, to impute under the next mot plausible assumption J2R, where the reference group is the standard care arm, we update the method specification to "j2r" and add that treatment=0 (standard care) is the reference group, using the `refgroup` option, as follows:

```
· mimix head treat, id(id) time(time) covariates(age sex migraine chronicity
  head_base) method(j2r) refgroup(0) m(50) regress clear seed(23) burnin(1000)
  burnbetween(500) saving(head_j2r0)
```

To impute under J2R where the reference arm is the acupuncture arm, we use the above code but alternatively indicate that refgroup=1. To impute under CIR, CR, or LMCF, the above line of code is adapted to include `cir`, `cr`, or `lmcf` in place of `j2r`, with the required reference group (or no reference group in the case of LMCF).

The primary MAR analysis (Table 3) suggests that acupuncture results in improved headache scores relative to standard care, with a treatment effect of −4.97. Sensitivity analysis results are sorted by plausibility (what we considered most plausible, down to what we considered least plausible). After MAR, we considered it most plausible that patients in the acupuncture arm discontinued treatment abruptly following their last observed outcome measurement and then jumped to follow the behavior observed in the standard care arm; unobserved outcomes of individuals in the standard care arm assumed to be MAR. Under jump to standard care, the treatment effect is lower at −3.32. Assuming that patients in the acupuncture arm more gradually tracked toward the standard care arm behavior following withdrawal, copy increments in standard care, the treatment effect is similar at −3.74. Then, if it is assumed that patients in the acupuncture arm with unobserved outcomes never undertook acupuncture and always behaved like the standard care patients, the treatment effect is slightly closer to MAR at −3.80. These three sensitivity analysis assumptions lead to a smaller treatment effect in comparison to MAR as the acupuncture patients with unobserved outcomes are assumed to have either copied the standard care profile for the entire trial duration or to have jumped to or copied the mean increments in the standard care arm (retaining their pre-drop out treatment arm means), which is itself higher (indicating poorer outcome).

The smallest treatment effect is obtained next under jump to acupuncture imputation, −3.00. The treatment effects are most extreme under the jump to reference assumptions since this assumes patients followed their own treatment arm means prior to deviation then abruptly switched to the mean profile of the alternative arm. Under jump to acupuncture, there are more patients abruptly switching behavior than under jump to standard care as more patients deviated in the standard care arm which is why jump to acupuncture sees the greatest difference in the treatment estimate, relative to the primary MAR analysis.

If the unobserved alternatively copied the acupuncture arm behavior throughout the trial, the treatment effect is −3.48. The results under copy increments in acupuncture are similar to this, −3.50. In these two cases, the standard care patients that dropped out are also assumed to behave like the acupuncture cases, either for the entire trial duration or to follow their mean increments post-drop out (retaining their pre-drop out treatment arm means). The assumptions that assume behavior of the acupuncture arm reduce the treatment difference by a larger amount relative to when standard care was the reference as there were more drop-outs in the standard care arm (see Figure 2); a greater proportion of patients are assumed to behave more similarly. Finally, under LMCF, which we considered the least plausible scenario, the treatment estimate is only very marginally lower than under MAR at −4.94. This corresponds to assuming the underlying mean response remains constant after drop out at the level of the patients randomized arm, that is, individuals in the acupuncture arm maintained the benefit of acupuncture at unobserved time points. Overall, under all scenarios, the treatment effect remains significant, indicating that we can be confident in the trials primary result that acupuncture is an effective treatment.

Reference-based MI can also be a useful tool for the sensitivity analysis of a trial with a single follow-up measurement. In Appendix C, we include a third case study to demonstrate the application of `mimix` for reference-based sensitivity analysis of clinical trials with a single follow-up.

5.2 | Incorporating different reasons for missingness

In the above sensitivity analyses, we have made the same assumption for the unobserved data for each patient with missing data. When reasons for withdrawal are available, we might prefer to vary the assumption for the missing data in the analysis according to the reason. In the acupuncture trial, additional data on the reason for withdraw were in fact recorded and are displayed in Table 4. Given these reasons, we might want to assume MAR for patients who withdrew as they experienced inter-current illness, died, or had an adverse effect and J2R (standard care as reference) for all others. The Stata command `mimix` enables different assumptions to be made for each deviating individual. To conduct this analysis, we first define a new variable which holds the required imputation method for each individual. Then we implement the MI analysis using the `methodvar` option with this newly created variable, in place of the previously used `method` using the following code.

```
· generate str4 method = "mar"
· replace method = "j2r" if withdrawal_reason == "treatment ineffective" | with-
  drawal_reason == "treatment hassle" | withdrawal_reason=="lost to follow-up" |
  withdrawal_reason== "withdrew consent"
· mimix head treat, id(id) time(time) covariates(age sex migraine chronicity
  head_base) methodvar(method) refgroup(0) m(50) regress clear seed(23) burnin(1000)
  burnbetween(500) saving(head_mar_j2r)
```

After creating the updated imputations the analysis model of interest is fitted to each imputed data set and results combined with Rubin's rules, using the `mi` estimate command. The treatment effect estimated by this sensitivity analysis is −3.74, SE 1.23, 95% CI −6.17 to −1.31, $p = 0.003$. Naturally, this lies between the estimate obtained under the assumption of MAR (TE = −4.97) and J2R for all unobserved data (TE = −3.32). But this may be considered a more realistic appropriate analysis within the acupuncture setting, given withdrawal reasons.

5.3 | Incorporating reference- and delta-based assumptions

In some situations, we might want to make a reference-based assumption for a specified group of patients (eg, J2R or CR) and a δ-based assumption for others. If a selection of patients require a δ-based adjustment, `mimix` can be run with a `methodvar`, which takes the value of MAR for these cases. The `methodvar` can take the value of the required reference

TABLE 4 Reasons for withdrawal in the acupuncture trial

Withdrawal reason	Standard care	Acupuncture	Total
Adverse effects	0	1	1
Died	1	0	1
Intercurrent illness	8	8	16
Lost to follow-up	7	8	15
Treatment hassle	0	5	5
Treatment ineffective	0	4	4
Withdrew consent	40	18	58
Total	56	44	100

based method(s) for other individuals as appropriate. Following execution of the MI, the δ-based adjustment can then be made as required to the MAR imputed data following the steps outlines in Section 4.

For example, suppose we wish to extend the above example and change the missing data assumption for the patients who withdrew due to intercurrent illness from MAR to be MAR+10, representing a worse headache score. We retain the assumption of MAR for the unobserved data of those who died or had an adverse effect and the assumption of J2R for those who withdrew due to all other reasons. We start by using the dataset imputed above (`head_mar_j2r`). We then implement the δ adjustment for the patients who withdrew due to intercurrent illness with the following code.

```
· use head_mar_j2r, clear
· mi passive: gen imputed=_mi_miss
· replace imputed=1 if imputed==.
· gen head_Delta1=head
· replace head_Delta1=head+10 if imputed==1 & withdrawal_reason=="intercurrent
  illness"
```

This gives a treatment effect of −3.74, SE 1.25, 95% CI −6.19 to −1.28, $p = 0.003$. This does not considerably change from the previous analysis. The point is `mimix` facilitates a wide variety of controlled MI analyses. We are not in fact limited to combining a δ adjustment following MAR imputation; it can be invoked following any reference based method. Countless possibilities exist!

6 | WHAT IS THE APPROPRIATE VARIANCE ESTIMATOR?

There has been debate around the appropriate variance estimator in the reference based MI settings.[48,49] This is because borrowing information between trial arms for MI produces peculiar behavior in the empirical long-run sampling variance of the reference based treatment effect. The strong assumption that deviators behave exactly like those observed in an opposing reference arm reduces the long-run sampling variance of the reference based treatment estimate, below that seen under MAR (typically the primary analysis assumption) and the variance that would be obtained, were the deviation data observed in the given reference based scenario. Also within the sensitivity analysis, the assumptions made at the imputation stage are not fully compatible with those of the primary analysis model, which is retained in the sensitivity analysis.[30]

The usual MI variance estimator, Rubin's MI variance estimator, exhibits entirely different behavior.[48,50,51] Rubin's variance estimator is always larger than the variance we would obtain had the deviation data been observed under the given scenario. Rubin's variance also increases as the proportion of missing data increases.[49] We have shown elsewhere[52] that Rubin's rules provide an appropriate estimate of variance for the treatment effect in reference based MI settings. Our justification being that Rubin's variance estimate provides *information anchored* inference. That is, the proportion of information lost due to missing data under MAR is approximately preserved in the sensitivity analysis. We regard information anchored inference as desirable. It ensures there is no loss or gain of information due to missing data in the sensitivity analysis relative to the primary analysis. Thus, regulators can be reassured the sensitivity analysis is not

injecting information, while trialists can be reassured that the sensitivity analysis is not discarding any of the valuable obtained data.

The results of the analyses presented here (Tables 3 and C2) demonstrate the information anchoring performance of Rubin's variance estimate. For both the headache trial and the reviewer trial, we obtain information anchored inference. We have also shown[52] that when employing δ-based MI with a fixed δ adjustment, information anchored inference will be obtained. We thus recommend the use of Rubin's variance estimator within delta- and reference-based sensitivity analyses.

6.1 | Incorporating a prior distribution on δ

Throughout Section 4, we assumed the parameter governing the difference between the MAR and MNAR distribution, δ, or parameters δ_m for missing data pattern m, are fixed. This does not have to be the case. The sensitivity analysis parameters may have their own distribution with specified mean and variance and thus vary over the imputation set K. For example, in the analysis of the ADAPT trial, δ was a fixed value to represent a postulated fixed difference in outcome between the observed and the unobserved, but we could have specified $\delta^k \sim N(\delta, \sigma^2_\delta)$ for imputation k. If δ has a distribution, this adds an additional step to the MI procedure where for each imputation k, δ^k must be drawn and all imputations for current imputation k edited by δ^k. This may be appropriate when there is uncertainty in the value of δ, when experts cannot come to a consensus on the likely difference in outcome between observed and unobserved cases. White et al[53] demonstrate successful elicitation of prior information on the difference between missing and observed outcomes in a single follow-up trial setting using a number of experts and a pre-specified questionnaire. They asked a number of experts what they thought δ would be using a standardized questionnaire then assessed the variability across the experts to form the prior variance on δ. Mason et al[54] present a practical tool for eliciting pooled expert opinion and demonstrate its use for randomized controlled trials with missing data.

What are the implications of incorporating an additional prior distribution on δ? We have shown elsewhere[52] that for a continuous normal outcome, where δ has an assumed normal prior, there will be a greater loss of information in the sensitivity analysis in comparison to the primary MAR analysis. That is, the variance of the MI treatment effect, estimated using Rubin's rules, will incorporate the additional variance on δ over k (because the variation in δ increases the between-imputation variance). In comparison to the primary analysis conducted under MAR, one will consequently obtain *information negative* inference, whereby there will be a greater loss of information due to missing data in the sensitivity analysis relative to the primary analysis.[52] By using a large enough variance for δ, we consequently have the ability to overturn the conclusions of any purportedly significant results reported in the primary analysis. Thus, when specifying δ, if incorporating a prior one must be confident they are using an appropriate value for the variance for δ. This should be determined in collaboration with experts in the field and will be context specific. It may be useful to vary the variance parameter on δ, as well as the specified mean value when the value of σ^2_δ is uncertain since this has implications for the variance of the treatment effect of interest. In order to avoid the loss of information in sensitivity analysis when δ is uncertain, a preferable option would be to conduct a number of fixed δ adjustments, with varying size for the fixed δ rather than employing a distribution on δ.

Another important to point to note is that, if the δ varies by missing data pattern m (or any alternative grouping) and the δ_ms are given a prior distribution, then the covariance of the δ_ms must also be specified for each missing data pattern m. It can be hard to elicit reliable information about this. Carpenter and Kenward[41] show that larger standard errors are obtained when the correlation is zero. This can, therefore, be a useful starting point. A subsequent complicating factor is that the variance of the final treatment estimate also then depends on the specified covariance of the δ_ms.

7 | DISCUSSION

In this tutorial paper, we have described and illustrated how sensitivity analysis can be conducted to explore departures from an MAR assumption for unobserved continuous outcome data using controlled MI. Controlled MI combines a pattern mixture modeling approach with MI. We have shown how we may adopt a controlled MI approach using numerical parameters which govern the degree of difference between the MAR and MNAR distributions, termed δ-based MI. We have also shown how specific sensitivity analysis parameter specification is not always required. One may alternatively

specify the difference between MAR and MNAR distributions qualitatively using entirely within trial information, by reference to other groups of individuals in the trial using reference-based MI.

The attraction of the different controlled MI approaches depends on the problem at hand. It is of course important that trialists employ only methods, which make plausible assumptions relevant to the clinical setting and estimand of interest. In some circumstances, the assumptions made by the reference based MI procedures may not be suitable. We advise trialists to consider carefully the assumptions behind the controlled MI analyses to ensure each analysis undertaken is suitable to the context at hand. Indeed, the assumptions behind any method of statistical analysis should be relevant to the clinical setting.

A different perspective and framing for the reference based MI options, jump to reference, CR, and CIR, is provided by White et al.[55] They show how these MI procedures are special cases of a causal model, which makes an explicit assumption about the maintained causal effect of treatment after discontinuation in a potential outcomes framework; under CIR, the treatment effect at the discontinuation visit is maintained, under CR, the treatment effect is assumed to be diminished, and under jump to reference, the treatment effect is assumed to be eliminated at the final visit. As such, additional reference-based MI procedures can be conceived where the maintained causal effect of treatment after discontinuation is varied in different ways.

We showed how δ-based MI can be implemented using the inbuilt MI package in Stata. Of course, one is not confined to using Stata and can use MI packages within other software to complete the analysis. We also showed how `mimix` can be used to conduct reference-based MI using the options described in Table 2 also within Stata. We also provided details of SAS programs that can be used if alternative software is preferred (at the time of writing, reference-based MI is in development for R). In Section 5, we noted that the reference based MI options described in Table 2 were not an exhaustive listing. Other controlled reference based MI procedures can be conceived. The `mimix` command implements only the five reference based options in Table 2. The "five macros" SAS package includes additional options, including "own last mean carried forward" (OLMCF), where the subject's own mean at withdrawal is carried forward. The "five macros" SAS package also enables user defined methods.

In the presence of missing data, when we do sensitivity analysis, we need to be sure we are not inappropriately injecting information or removing information within the analysis regardless of the specific method utilized. As discussed within, it has been shown elsewhere that the "δ method" of MI with a fixed δ adjustment and reference based MI both preserve the information loss observed under MAR;[52] information anchored inference is obtained. This provides relevant, accessible, justified inference in the context of missing data sensitivity analysis. Trialists can be confident when utilizing these approaches that they are not unnecessary loosing or gaining any information beyond that observed under MAR. Care should, however, be exercised when using the "δ-method" with a prior distribution on δ to ensure an appropriate variance is used for δ since the variance of the resulting treatment estimate will incorporate the variance on δ.

We have focused throughout on the analysis of a continuous outcome. However implementations of delta- and reference-based MI procedures do exist for other types of outcome variables. For binary outcomes, delta-based MI is available in Stata using the `mi impute logit` command with the `offset` option. To conduct reference based analysis, Rehal et al[56] propose modeling the binary data as if it was continuous and employing the methods illustrated in this tutorial. Following imputation, missing observations imputed as continuous are back-transformed to binary observations using an adaptive rounding algorithm.

For recurrent event data with discrete time points, Keene et al[57] have applied the negative binomial distribution for reference-based imputation. The event rate in a specified reference arm is used to impute missing data in other arms of a trial. A SAS macro implementing their approach, written by James Roger, and an R package `dejaVu`, is available at http://www.missingdata.org.uk. Akache and Ogundimu[58] extended the approach of Keene et al by modeling the recurrent event data process in continuous time. Gao et al[59] alternatively described how a piecewise exponential model can be used to model a recurrent event outcome to implement reference based MI.

For a survival outcome, Jackson et al[60] showed how one can allow for non-independent censoring in a Cox proportional hazard model. Their method first fits a Cox model to the observed data under the usual (conditional on covariates) noninformative censoring assumption. Multiple imputed datasets are then generated assuming that the hazard for failure following censoring changes by a user specified multiplier (a hazard ratio) compared with the hazard implied by the non-informative censoring assumption. This can be implemented using the R package `InformativeCensoring`. For survival data, Atkinson has proposed a collection of reference-based assumptions using a proportional hazards model.[61,62] Lu et al[63] compared a δ adjusted MI method and a reference based MI method for survival data. The δ adjusted method evaluated specifies that the hazard of having the event of interest for subjects who discontinued before the time point of interest is multiplicatively increased relative to that for those who continued in the trial beyond the time point of interest.

The reference based method imputes assuming the hazard for a patient who discontinued lies between the hazard for those who continued and the hazard for the reference group. Lipkovich[64] also proposed a δ-based approach for a survival outcome using a variety of imputation models, including a semi-parametric Cox model, piecewise exponential baseline hazard function, Kaplan-Meier method, and longitudinal logistic regression for event occurrence in discretized intervals of time. Most recently, Tang[65] proposed two approaches for implementing controlled imputation (including both reference and delta based MI) for binary and ordinal data based respectively on the sequential logistic regression and multivariate probit model.

We have assumed throughout the reference based analyses that baseline covariates are complete. When `mimix` is used to conduct reference-based MI, any included covariates must be complete. These can be filled in using an appropriate method prior to implementing `mimix` if this is not the case. For example, in randomized trials missing baseline covariates may be imputed using the mean value for all cases with non-missing data imputation. This is justifiable, since randomization ensures that baseline variables are independent of treatment group.[66] When δ based MI is used, a similar approach may be utilized, or the baseline covariates may be included in the initial MAR imputation model for imputation. Additionally, a potential limitation of the reference-based approach is that its implementation relies on the assumption of an underlying multivariate-normal model. However, Shafer[35] reported that imputations drawn under the MVN are robust to skewness, as long as the estimators of interest are normally distributed. Under MAR MI may be implemented via chained equations to avoid such assumptions (MICE).

Of course, many other methods, all characterizable as non-multiple-imputation, are available for conducting sensitivity analysis of clinical trials with missing data.[4-6] We advocate the outlined principled controlled MI route since, as we have found in our practical trial experience and, as we hope to have demonstrated here, it is a pleasingly accessible option. No direct estimation of an MNAR model is required, which can often require complex one-off coding. Both controlled MI options enable open and interpretable sensitivity analysis, with assumptions that trial personnel can understand. We have also shown how, in a single trial setting, both delta- and reference-based assumptions may be made to allow for different types of missingness.

A wealth of sensitivity analyses are possible with the described and illustrated methods.

ACKNOWLEDGMENTS

We thank Dr. Susan Chan and the ADAPT trial team for use of data from the ADAPT study within this tutorial. We thank Dr. Andrew Vickers and the acupuncture trial team for the use of their data within this tutorial. We would also like to thank Dr. Sara Schroter and the reviewer trial team for the their data. Additionally we thank Ping-tee Tan for her helpful comments on the manuscript. James R Carpenter and Tim P Morris are supported by the Medical Research Council (grant numbers MC_UU_12023/21 and MC_UU_12023/29).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

AUTHOR CONTRIBUTIONS

SC, JC, and MK conceived this tutorial. SC conducted the analyses within this tutorial and drafted this manuscript. SC, JC, TP, and MK all interpreted the data and critically reviewed the manuscript. All authors read and approved the final manuscript.

ORCID

Suzie Cro https://orcid.org/0000-0002-6113-1173
Tim P. Morris https://orcid.org/0000-0001-5850-3610
James R. Carpenter https://orcid.org/0000-0003-3890-6206

REFERENCES

1. Morris TP, Kahan BC, White IR. Choosing sensitivity analyses for randomised trials: principles. *BMC Med Res Methodol*. 2014;14(1):11. https://doi.org/10.1186/1471-2288-14-11.
2. Committee for Medicinal Products for Human Use. *Guideline on Missing Data in Confirmatory Clinical Trials*. London: Eurpoean Medicines Agency; 2010. https://www.ema.europa.eu/en/missing-data-confirmatory-clinical-trials.
3. National Research Council. The prevention and treatment of missing data in clinical trials. panel on handling missing data in clinical trials. *Committee on National Statistics, Division of Behavioural and Social Sciences Education*. Washington, DC: The National Academies Press; 2010.

4. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. New York, NY: Wiley; 2007.
5. Carpenter JR, Kenward MG. *Multiple Imputation and its Application*. New York, NY: Wiley; 2013.
6. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC Handbooks of Modern Statistical MethodsTaylor & Francis; 2014.
7. Bell ML, Fiero M, Horton NJ, Chiu-Hsieh H. Handling missing data in RCTs, a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):118.
8. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials*. 2014;15:237. https://doi.org/10.1186/1745-6215-15-237.
9. Ich, C. H. M. P. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Proceedings of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use; 2019.
10. Ratitch B. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Stat*. 2013;12(6):337-347. https://doi.org/10.1002/pst.1549.
11. Carpenter JR, Roger JH, Kenward MG. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions and inference via multiple imputation. *J Biopharmaceut Stat*. 2013;23(6):1352-1371.
12. Mallinckrodt C, Roger J, Chuang-Stein C, et al. Recent Developments in the Prevention and Treatment of Missing Data. *Therapeut Innovat Regulat Sci*. 2014;48(1):68-80. https://doi.org/10.1177/2168479013501310.
13. Mallinckrodt CH. Preventing and treating missing data in longitudinal clinical trials: a practical guide. *Practical Guides to Biostatistics and Epidemiology*. Cambridge, MA: Cambridge University Press; 2013.
14. O'Kelly M, Ratitch B. *Multiple Imputation*. New York, NY: John Wiley & Sons Ltd; 2014:284-319.
15. Ayele BT, Lipkovich I, Molenberghs G, Mallinckrodt CH. A multiple-imputation-based approach to sensitivity analyses and effectiveness assessments in longitudinal clinical trials. *J Biopharmaceut Stat*. 2014;24(2):211-228.
16. Kenward MG. Controlled multiple imputation methods for sensitivity analyses in longitudinal clinical trials with dropout and protocol deviation. *Clin Investigat*. 2015;5(3):311-320.
17. Philipsen A, Jans T, Graf E, et al. Effects of group psychotherapy, individual counseling, methylphenidate, and placebo in the treatment of adult attention-deficit/hyperactivity disorder: a randomized clinical trial. *JAMA Psychiat*. 2015;72(12):1199-1210. https://doi.org/10.1001/jamapsychiatry.2015.2146.
18. Jans T, Jacob C, Warnke A, et al. Does intensive multimodal treatment for maternal ADHD improve the efficacy of parent training for children with ADHD? a randomized controlled multicenter trial. *J Child Psychol Psychiatry*. 2015;56(12):1298-1313. https://doi.org/10.1111/jcpp.12443.
19. Billings LK, Doshi A, Gouet D, et al. Efficacy and safety of IDegLira versus basal-bolus insulin therapy in patients with type 2 diabetes uncontrolled on metformin and basal insulin; DUAL VII randomized clinical trial. *Diabet Care*. 2018;41(5):1009-1016. https://doi.org/10.2337/dc17-1114.
20. Atri A, Frolich L, Ballard C, et al. Effect of idalopirdine as adjunct to cholinesterase inhibitors on change in cognition in patients with Alzheimer disease: three randomized clinical trials. *JAMA*. 2018;319(2):130-142. https://doi.org/10.1001/jama.2017.20373.
21. Chan S, Cornelius VR, Chen T, et al. Atopic Dermatitis Anti-IgE Paediatric Trial (ADAPT): the role of anti-IgE in severe Paediatric Eczema: study protocol for a randomised controlled trial. *Trials*. 2017;18(1):136. https://doi.org/10.1186/s13063-017-1809-7.
22. Chen T, Chan S, Lack G, Cro S, Cornelius VR. The role of anti-IgE (omalizumab/Xolair) in the management of severe recalcitrant paediatric atopic eczema (ADAPT): statistical analysis plan. *Trials*. 2017;18(1):231. https://doi.org/10.1186/s13063-017-1976-6.
23. Chan S, Cornelius V, Cro S, Harper JI, Lack G. Treatment Effect of Omalizumab on Severe Pediatric Atopic Dermatitis: The ADAPT Randomized Clinical Trial. *JAMA Pediatr*. 2020;174(1):29-37. https://doi.org/10.1001/jamapediatrics.2019.4476.
24. Vickers AJ, Rees RW, Zollman CE, et al. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *BMJ*. 2004;328(7442):744. https://doi.org/10.1136/bmj.38029.421863.EB.
25. Vickers AJ. Whose data set is it anyway? sharing raw data from randomized trials. *Trials*. 2006;7:15-15. https://doi.org/10.1186/1745-6215-7-15.
26. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.
27. Wilson E, Free C, Morris TP, et al. Internet-accessed sexually transmitted infection (e-STI) testing and results service: a randomised, single-blind, controlled trial. *PLOS Med*. 2017;14(12):1-20. https://doi.org/10.1371/journal.pmed.1002479.
28. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Son, Inc; 1987.
29. Rubin DB. *Multiple Imputation for Nonreponse in Surveys*. Wiley Series in Probability and Mathematical Statistics. New York, NY: John Wiley & Sons; 1987.
30. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9(4):538-558.
31. Royston P, Carlin JB, White IR. Multiple imputation of missing values: new features for mim. *Stat J*. 2009;9(2):252-264.
32. Rubin DB. Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. Paper presented at: Proceedings of the Survey Research Methods Section of the American Statistical Association; 1978:20-28.
33. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc*. 1996;91(434):473-489.
34. Little RJA, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics*. 1996;52(4):1324-1333.
35. Schafer JL. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall; 1997.
36. Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*. 1999;18(6):681-694. https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.

37. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. https://doi.org/10.1002/sim.4067.
38. Hippel PT. How many imputations do you need? a two-stage calculation using a quadratic rule. *Sociolog Methods Res*. 2018;1(1):1-20. https://doi.org/10.1177/0049124117747303.
39. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res*. 2007;16(3):219-242. https://doi.org/10.1177/0962280206074463.
40. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc*. 1993;88(421):125-134.
41. Carpenter JR, Kenward MG. Missing data in randomised controlled trials - a practical guide; 2008. http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf. Accessed June 2014.
42. Daniels MJ, Hogan JW. *Missing Data in Longitudinal Studies Strategies for Bayesian Modelling and Sensitivity Analysis*. Monographs on Statistics and Applied Probability. Boca Raton, FL: Chapman & Hall; 2008.
43. White IR, Horton NJ, Carpenter JR, Pocock SJ. Strategy for intention to treat analysis in randomised trials with missing outcome data. *British Med J*. 2011;342:d40.
44. Yan X, Lee S, Li N. Missing data handling methods in medical device clinical trials. *J Biopharmaceut Stat*. 2009;19(6):1085-1098.
45. Cro S, Morris TP, Kenward MG, Carpenter JR. Reference-based sensitivity analysis via multiple imputation for longitudinal trials with protocol deviation. *Stat J*. 2016;16(2):443-463(21).
46. Roger JH. miwithd; SAS code for reference based multiple imputation; 2012. www.missingdata.org.uk. Accessed July 15 2016.
47. Roger JH., Barnett C., Drury T. The five macros; SAS code for reference based multiple imputation; 2017. www.missingdata.org.uk. Accessed February 16, 2018.
48. Seaman SR, White IR, Leacy FP. Comment on, analysis of longitudinal trials with protocol deviations, a framework for relevant, accessible assumptions, and inference via multiple imputation, by Carpenter, Roger and Kenward. *J Biopharmaceut Stat*. 2014;24(6):1358-1362.
49. Carpenter JR, Roger JH, Cro S, Kenward MG. Response to comments by seaman et al. on analysis of longitudinal trials with protocol deviation, a framework for relevant, accessible assumptions, and inference via multiple imputation. *J Biopharmaceut Stat*. 2014;24(6):1363-1369.
50. Lu K. An analytic method for the placebo-based pattern- mixture model. *Stat Med*. 2014;33(7):1134-1145. https://doi.org/10.1002/sim.6008.
51. Ayele BT, Lipkovich I, Molenberghs G, Mallinckrodt CH. A multiple-imputation-based approach to sensitivity analyses and effectiveness assessments in longitudinal clinical trials. *J Biopharmaceut Stat*. 2014;24(2):211-228. https://doi.org/10.1080/10543406.2013.859148.
52. Cro S, Carpenter JR, Kenward MG. Information-anchored sensitivity analysis: theory and application. *J Royal Stat Soc Ser A (Stat Soc)*. 2019;182(3):623-645.
53. White IR, Carpenter JR, Evans S, Schroter S. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clin Trials*. 2007;4:125-139.
54. Mason Alexina J, Gomes Manuel, Grieve Richard, Ulug Pinar, Powell Janet T, Carpenter James. Development of a practical approach to expert elicitation for randomised controlled trials with missing health outcomes: Application to the IMPROVE trial. *Clin Trials* 2017;14(4):357-367. PMID: 28675302, https://doi.org/10.1177/1740774517711442.
55. White Ian R., Joseph Royes, Best Nicky. A causal modelling framework for reference-based imputation and tipping point analysis in clinical trials with quantitative outcome. *J Biopharmaceut Stat* 2020;30(2):334-350. PMID: 31718423, https://doi.org/10.1080/10543406.2019.1684308.
56. Rehal S. Implications of Missing Data in Tuberculosis Non-inferiority Clinical Trials (PhD Thesis). London, UK: University College London, University of London; 2018.
57. Keene ON, Roger JH, Hartley BF, Kenward MG. Missing data sensitivity analysis for recurrent event data using controlled imputation. *Pharm Stat*. 2014;13(4):258-264. https://doi.org/10.1002/pst.1624.
58. Akacha M, Ogundimu EO. Sensitivity analyses for partially observed recurrent event data. *Pharmaceut Stat*. 2015;15(1):4-14. https://doi.org/10.1002/pst.1720.
59. Gao F, Liu GF, Zeng D, et al. Control-based imputation for sensitivity analyses in informative censoring for recurrent event data. *Pharm Stat*. 2017;16(6):424-432. https://doi.org/10.1002/pst.1821.
60. Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Stat Med*. 2014;33(27):4681-4694. https://doi.org/10.1002/sim.6274.
61. Atkinson A. Reference Based Sensitivity Analysis for Time-to-Event Data (PhD thesis). London, UK: Department of Medical Statistics, London School of Hygiene & Tropical Medicine, University of London; 2018.
62. Atkinson A, Kenward MG, Clayton T, Carpenter JR. Reference-based sensitivity analysis for time-to-event data. *Pharmaceut Stat*. 2019;18:645-658. https://doi.org/10.1002/pst.1954.
63. Lu K, Li D, Koch GG. Comparison between two controlled multiple imputation methods for sensitivity analyses of time-to-event data with possibly informative censoring. *Stat Biopharmaceut Res*. 2015;7(3):199-213. https://doi.org/10.1080/19466315.2015.1053572.
64. Lipkovich I, Ratitch B, O'Kelly M. Sensitivity to censored-at-random assumption in the analysis of time-to-event endpoints. *Pharmaceut Stat*. 2016;15(3):216-229. https://doi.org/10.1002/pst.1738.
65. Tang Y. Controlled pattern imputation for sensitivity analysis of longitudinal binary and ordinal outcomes with nonignorable dropout. *Stat Med*. 2018;37(9):1467-1481. https://doi.org/10.1002/sim.7583.
66. White IR, Thompson SG. Adjusting for partially missing baseline measurements in randomized trials. *Stat Med*. 2005;24(7):993-1007. https://doi.org/10.1002/sim.1981.

67. Schroter S, Black N, Evans S, Carpenter JR, Fiona G, Smith R. Effects of training on quality of peer review: randomised controlled trial. *BMJ*. 2004;328(7441):673. https://doi.org/10.1136/bmj.38023.700775.AE.

How to cite this article: Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*. 2020;39:2815–2842. https://doi.org/10.1002/sim.8569

APPENDIX A. Δ-BASED MI INCORPORATING REASONS FOR MISSINGNESS AND TIME OF DROP-OUT

Using the ADAPT dataset, described in Section 3.4, we present code to conduct δ-based MI, which incorporates (i) the reasons for missing data and (ii) the time of drop out using δ as a change in rate of response. A third sensitivity analysis is presented white incorporates (i) and (ii). To do so, here we make use of two additional variables in the ADAPT dataset, `reason` and `d_time`. `reason` indicates the reason for missing data as `reason=1` for patients who withdrew from treatment and `reason=2` for patients who received rescue medication. `d_time` provides the last observed time point for each patient in weeks as 0, 4, 8, 12, 20, or 24.

1. δ-based MI using a fixed delta adjustment at the last time point (week 24) which varies by reason for missing data. δ=7.5 for individuals who withdrew from treatment. δ=3.75 for individuals who received rescue therapy.
```
· use adapt_MAR, clear
· mi passive: gen imputed=_mi_miss
· replace imputed=1 if imputed==.
· gen CDLQI_w24_Delta_A1=CDLQI_w24
· replace CDLQI_w24_Delta_A1=CDLQI_w24 + 7.5 if imputed==1 & reason==1
· replace CDLQI_w24_Delta_A1=CDLQI_w24 + 3.75 if imputed==1 & reason==2
· mi estimate : regress CDLQI_w24_Delta_A1 i.treat i.agestrat CDLQI_1 i.IgEstrat
```
2. δ-based MI using δ as a fixed change in rate of 1.25 in (C)DLQI decline for every 4 weeks unobserved. 1.25 corresponds to 100% of the absolute rate of change in (C)DLQI seen in the trial over 4 weeks assuming a constant linear rate of change. Here δ does not vary by reason for missing data.
```
· gen CDLQI_w24_Delta_A2=CDLQI_w24
· replace CDLQI_w24_Delta_A2=CDLQI_w24 + 1.25(24-d_time)/4 if imputed==1
· mi estimate : regress CDLQI_w24_Delta_A2 i.treat i.agestrat CDLQI_1 i.IgEstrat
```
3. δ-based MI using δ as a fixed change in rate of (C)DLQI decline which varies by missing data pattern. δ=1.25 for every 4 weeks unobserved for individuals who withdrew from treatment. δ=0.625 for every 4 weeks unobserved for individuals who received rescue therapy.
```
· gen CDLQI_w24_Delta_A3=CDLQI_w24
· replace CDLQI_w24_Delta_A3=CDLQI_w24 + 1.25(24-d_time)/4 if imputed==1 & rea-
  son==1
· replace CDLQI_w24_Delta_A3=CDLQI_w24 + 0.625(24-d_time)/4 if imputed==1 &
  reason==2
· mi estimate : regress CDLQI_w24_Delta_A3 i.treat i.agestrat CDLQI_1 i.IgEstrat
```

APPENDIX B. TECHNICAL DETAILS OF REFERENCE-BASED MI

Reference-based MI was originally introduced by Little and Yau in 1996,[34] who imputed unobserved data for patients under the dose of treatment actually received rather than as assigned using monotone regression. More recently in 2013, Carpenter et al[11] formalized the approach and presented a novel collection of five MI procedures for reference based sensitivity analysis which we focus on here. The generic reference based MI algorithm of Carpenter et al,[11] for longitudinal trials with a continuous outcome, is now presented (with minor modifications):

1. Separately for each treatment arm take all the observed data, and assuming MAR, fit a MVN distribution with an unstructured mean (ie, a separate mean for each of the baseline and post-randomization observation times) and variance-covariance matrix using a Bayesian approach with an improper prior for the mean and an uninformative Jeffrey's prior for the covariance matrix.
2. Draw a mean vector and covariance matrix from the posterior distribution for each treatment arm. Specifically we use the MCMC method to draw from the appropriate Bayesian posterior, with a sufficient burn-in and update the chain sufficiently in-between to ensure subsequent draws are independent, given the observed data. The sampler can be initiated using the EM algorithm.
3. Use the draws in step 2 to form the joint distribution of each deviating individual's observed and missing outcome data as required. This can be done under a range of assumptions, in order to explore the robustness of inference about the treatment effects. The options presented by Carpenter et al[11] that each translate to a relevant assumption are described in Table 2.
4. Construct the conditional distribution of missing given observed outcome data for each deviating individual, using their joint distribution formed in step 3. Sample missing data from the conditional distributions to create a completed dataset.
5. Repeat steps 2 to 4 K times, resulting in K imputed datasets.

Here we describe how the algorithm proceeds for the headache trial, to demonstrate what is going on behind the scenes when `mimix` is used. Initially a MVN distribution with an unformative prior is fitted to the observed data, assuming MAR, with an unstructured mean and variance-covariance matrix separately by treatment arm. A draw of the MAR mean vector and variance-covariance matrix by treatment arm is obtained via the MCMC method from the fitted Bayesian posterior (following sufficient burn in). We denote the current draw of the acupuncture group means and variance-covariance by $\boldsymbol{\mu}_a^k = \left[\mu_{a0}^k, \mu_{a3}^k, \mu_{a12}^k\right]$ and $\boldsymbol{\Sigma}_a^k$. The current draw of the standard care group means and variance-covariance from the posterior is denoted by $\boldsymbol{\mu}_c^k = \left[\mu_{c0}^k, \mu_{c3}^k, \mu_{c12}^k\right]$ and $\boldsymbol{\Sigma}_c^k$. Figure 4A is a schematic illustration of a current draw of the MAR means. The two solid triangles in Figure 4 represent observations from a randomly selected acupuncture patient who was lost to follow-up some time following month 3 and has unobserved data at month 12.

Under the primary MAR assumption, the joint distribution of a patient is observed and missing data are formed as MVN with mean and variance-covariance matrix from their randomized arm. For the randomly selected active patient who was lost to follow-up in our example, their joint data distribution for imputation k is formed as MVN with mean $\boldsymbol{\mu}_a^k$ and covariance matrix $\boldsymbol{\Sigma}_a^k$, as shown by the dotted line in Figure 4B.

In sensitivity analysis, under J2R where the standard care arm is the reference, the joint distribution of a patient's observed and missing data under J2R for imputation k is formed as MVN with mean $\boldsymbol{\mu}_i^k = \left[\mu_{a0}^k, \mu_{a3}^k, \mu_{c12}^k\right]$ as illustrated in Figure 4C. The proposed variance-covariance matrix for patient i for imputation k,

$$\boldsymbol{\Sigma}_{i,J2R}^k = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^k & \boldsymbol{\Sigma}_{12}^k \\ \boldsymbol{\Sigma}_{21}^k & \boldsymbol{\Sigma}_{22}^k \end{bmatrix},$$

is constructed by first partitioning the current posterior draws of the acupuncture and standard care variance-covariance matrices by patient i's observed and missing measurements. Below $\boldsymbol{\Sigma}_a^k$ and $\boldsymbol{\Sigma}_c^k$ have been accordingly partitioned; 1 indexes observed measurements (baseline and month 3) and 2 indexes missing measurements (month 12).

$$\boldsymbol{\Sigma}_a^k = \begin{bmatrix} \mathbf{a}_{11}^k & \mathbf{a}_{12}^k \\ \mathbf{a}_{21}^k & \mathbf{a}_{22}^k \end{bmatrix},$$

$$\boldsymbol{\Sigma}_c^k = \begin{bmatrix} \mathbf{c}_{11}^k & \mathbf{c}_{12}^k \\ \mathbf{c}_{21}^k & \mathbf{c}_{22}^k \end{bmatrix}.$$

Then as shown by Carpenter et al,[11] $\boldsymbol{\Sigma}_{11}^k = \mathbf{a}_{11}^k$, $\boldsymbol{\Sigma}_{12}^k = \boldsymbol{\Sigma}_{21}^k = \mathbf{p}_{21}^k \left(\mathbf{p}_{11}^k\right)^{-1} \mathbf{a}_{11}$ and $\boldsymbol{\Sigma}_{22} = \mathbf{p}_{22} - \mathbf{p}_{21}\left(\mathbf{p}_{11}^k\right)^{-1}\left(\mathbf{p}_{11} - \mathbf{a}_{11}\right)\left(\mathbf{p}_{11}^k\right)^{-1}\mathbf{p}_{12}$.

Under CIR, $\boldsymbol{\mu}_i^k = \left[\mu_{a0}^k, \mu_{a3}^k, \mu_{a3}^k + (\mu_{c12}^k - \mu_{c3}^k)\right]$, as depicted in Figure 4D. The variance-covariance matches the randomized arm for observed measurements and the standard care arm for the conditional components of missing given

TABLE C1 Peer review study: quality of peer review at baseline for those who did and did not complete the second review

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

observed measurements. That is, for imputation k, $\boldsymbol{\Sigma}^k_{i,CIR} = \boldsymbol{\Sigma}^k_{i,J2R}$. For any individuals already in the reference group this means, like J2R, their missing data will be imputed under MAR.

Under the CR, as shown in Figure 4D, $\boldsymbol{\mu}^k_i = \left[\mu^k_{c0}, \mu^k_{c3}, \mu^k_{c12}\right]$. The variance-covariance matrix is $\boldsymbol{\Sigma}^k_{i,CR} = \boldsymbol{\Sigma}^k_p$. Finally, under LMCF, $\boldsymbol{\mu}^k_i = \left[\mu^k_{a0}, \mu^k_{a3}, \mu^k_{a3}\right]$ as illustrated in Figure 4E. The variance-covariance matrix matches the randomized arm, that is, $\boldsymbol{\Sigma}^k_{i,LMCF} = \boldsymbol{\Sigma}^k_a$.

Under all the alternative reference based assumptions, after forming the required joint distributions for each patient with missing data, we construct the appropriate conditional distributions and the pieced together joint distributions imply for the missing data, given the observed data. A random sample is drawn from the formed conditional distributions to complete the patients' unobserved measurements for imputation k. Subsequently a new draw of the MAR mean vectors and variance-covariance matrices by treatment arm, via the MCMC method from the fitted Bayesian posterior, is retained (after sufficient burn between) and the formation of the required distributions and drawing of missing data is repeated, for the number of imputations required.

APPENDIX C. ANALYSIS OF A TRIAL WITH ONLY ONE FOLLOW-UP

C.1 Peer review trial

In this section, we introduce a case study to demonstrate the practical application of reference based MI for sensitivity analysis of clinical trials with a single follow-up. The data come from a randomized controlled trial, evaluating the impact of training on the quality of peer review conducted by Schroter et al.[66] In the original trial, 609 participants were randomized to receive either no-training, face-to-face training, or a self-taught package. Each participant was sent a baseline paper to review (paper 1) and the review quality was measured by two blinded researchers using the Review Quality Index (RQI) which results in a score ranging from 1 (worst) to 5 (best). Two to three months later, participants who had completed their first review were sent a further article to review (paper 2) and the RQI was measured; if this was returned, a third paper was sent three months later (paper 3) and again the RQI was measured. Unfortunately not all of the reviewers returned the required reviews. The original trial analysis was conducted under the MAR assumption, using a linear regression of the RQI on treatment group adjusted for baseline RQI. The analysis showed that the review quality of paper 2 was significantly higher for the self-taught group in comparison to the no-training group. Table C1 shows the quality of the review at baseline for (i) those who went on to complete the second review and (ii) those who did not, for each of these two interventions. The results suggest that a disproportionate number of poor reviewers in the self-taught group failed to review paper 2. We focus here on examining the robustness of this purportedly significant result. We will use reference based MI to establish what the results would look like if we assume that the reviewers who did not return paper 2 behaved like those in the no-training group (essentially assuming that the unobserved had not undertaken the self-taught training).

C.2 Analysis

Within the reviewer dataset, `id` is the unique reviewer identification number, `inter` indicates the randomly assigned intervention package of no training (`inter` = 0) or self-taught package (`inter` = 1), `base` is the baseline mean review quality, and `resp` is the mean review quality response for paper 2. The original primary analysis (complete case MAR) followed by an MAR MI analysis and CR MI analysis, where the reference arm is the no training arm, will be conducted. For each of the sensitivity analyses, 50 imputations will be produced with a burn in of 1000 and burn between of 500. The analysis model is a linear regression of review quality of paper 2 on treatment group, adjusted for baseline RQI as in the original primary analysis. The original complete case primary analysis of the trial followed by a standard MI analysis under MAR using `mimix` is first conducted using the below code. `mimix` requires the data to include a time indicator, which in this simple setting is not relevant so we first create a dummy variable to represent time, prior to calling

TABLE C2 Sensitivity analysis results for the reviewer study

Analysis	Treatment Est.	95%CI	Std. Err.	*P*-value
Complete case (MAR)	0.237	0.099 to 0.376	0.070	0.001
MI, "on-treatment MAR"	0.248	0.112 to 0.384	0.069	<0.001
MI, copy no training	0.177	0.043 to 0.311	0.068	0.010

reference based MI under CR, where the reference group is the no training treatment group. The commands required are as follows.

```
· use reviewer, clear
· regress resp inter base
· generate byte time = 2
· mimix resp inter, id(id) time(time) covariates(base) method(mar) m(50) regress
  burnin(1000) burnbetween(500) clear seed(23)
```

We now use `mimix` to establish the treatment effect if the reviewers who did not return paper 2 copied the no training arm behavior (`inter`= 0). The commands required are as follows.

```
· use reviewer, clear
· gen time = 2
· mimix resp inter, id(id) time(time) covariates(base) method(cr) refgroup(0) m(50)
  regress burnin(1000) burnbetween(500) clear seed(23)
```

Results from the analyses of the peer review study are summarized in Table C2. We see the intervention effect under copy no training behavior is slightly reduced at 0.177, compared with 0.237 from complete case analysis and 0.248 from standard MI analysis (both under MAR), but it remains statistically significant.