Taylor & Francis
Taylor & Francis Group

Reference-Based Multiple Imputation—What is the Right Variance and How to Estimate It

Jonathan W. Bartlett[a]

Department of Mathematical Sciences, University of Bath, Bath, UK

ABSTRACT

Reference-based multiple imputation methods have become popular for handling missing data in randomized clinical trials. Rubin's variance estimator is well known to be biased compared to the reference-based imputation estimator's true repeated sampling (frequentist) variance. Somewhat surprisingly given the increasing popularity of these methods, there has been relatively little debate in the literature as to whether Rubin's variance estimator or alternative (smaller) variance estimators targeting the repeated sampling variance are more appropriate. We review the arguments made on both sides of this debate, and argue that the repeated sampling variance is more appropriate. We review different approaches for estimating the frequentist variance, and suggest a recent proposal for combining bootstrapping with multiple imputation as a widely applicable general solution. At the same time, in light of the consequences of reference-based assumptions for frequentist variance, we believe further scrutiny of these methods is warranted to determine whether the strength of their assumptions is generally justifiable.

1. Introduction

Reference or reference-based imputation approaches have become popular for handling missing data in randomized trials (Carpenter, Roger, and Kenward 2013). They have typically been used as a sensitivity analysis to a primary analysis which assumes data are missing at random (MAR). Unlike most missing not at random (MNAR) sensitivity analysis methods, which often require specification of the value of sensitivity parameters, reference-based methods make assumptions which can be described somewhat more qualitatively by specifying the distribution of missing data in the active arm by reference to the distribution in the reference or control arm. In the case of a continuous outcome which is measured repeatedly over time, popular reference-based approaches include Jump to reference (J2R) and copy reference (CR). In J2R, as described in further detail in Section 2, the imputation distribution is constructed by assuming for active treatment group patients with missing data that marginally their visit-specific means before dropout are equal to the active group means and after dropout are equal to the reference group (e.g., control group) visit-specific means. In CR, all means are taken from the reference group. Thus, broadly speaking, reference-based methods impute post dropout outcomes for patients in the active group as if they switched at the point of dropout to the control treatment.

Although reference-based MI has since its inception tended to be used for sensitivity analyses for missing data, its core idea has been adopted more recently to handle intercurrent events for estimation of certain estimands (Mallinckrodt et al. 2019). Moreover, a recent review of trials published in the

Lancet and *New England Journal of Medicine* between 2014 and 2019 showed that the use of reference or control based MI is increasing in clinical trials (Tan et al. 2021).

An important question when using reference-based MI is how to estimate the variance of the resulting estimator. Conventionally the variance of MI estimators is obtained using Rubin's combination rules. However, as observed by Seaman, White, and Leacy (2014), for reference-based MI Rubin's variance estimator is biased upward relative to the repeated sampling variance of the MI estimator. This bias is due to uncongeniality between the imputation and analysis models (Meng 1994). Subsequently, a number of authors have discussed the merits of using either Rubin's variance estimator (Carpenter et al. 2014; Cro, Carpenter, and Kenward 2019) or the frequentist variance (Seaman, White, and Leacy 2014; Lu 2014; Tang 2017; White, Joseph and Best 2020), but no consensus has emerged. As we describe in more detail in Section 4, the frequentist variance can be estimated by analytical approaches such as the delta method (e.g., as described by Tang 2017) or by bootstrapping (e.g., as described by Gao, Liu, Zeng, Diao, Heyse, and Ibrahim 2017).

In Section 2, we review reference-based MI methods for continuous endpoints, the definition of congeniality, and why Rubin's variance estimator is biased upward relative to the frequentist variance of the reference-based point estimator of treatment effect. In Section 3, we review arguments made in favor of both Rubin's variance estimator and the frequentist variance, concluding that if the assumptions made by reference-based MI are employed, the frequentist variance is the right one. In Section 4, we review different approaches for estimating the frequentist variance of reference-based estimators. In Section 5,

we report the results of a simulation study for a recurrent event endpoint, and in Section 6 give some discussion.

2. Reference-Based Multiple Imputation and Congeniality

In this section we review reference-based multiple imputation methods and the congeniality issue. The approach was originally proposed in the context of a repeatedly measured continuous endpoint assuming a multivariate normal model (Carpenter, Roger, and Kenward 2013). Subsequently, the idea has been extended to other endpoint types, including recurrent events (Keene et al. 2014) and survival times (Atkinson et al. 2019). To help make the following arguments regarding congeniality concrete yet (relatively) simple, we first review the J2R approach for a repeatedly measured continuous endpoint, following Carpenter, Roger, and Kenward (2013).

2.1. J2R Imputation

We assume that $i = 1, \ldots, n$ patients are randomized to either reference (denoted $X_i = 0$) or active treatment ($X_i = 1$). Patients are scheduled to have the outcome measured at times $j = 0, \ldots, J$. The outcomes intended to be measured are thus $\mathbf{Y}_i = (Y_{i0}, Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$. Often there may exist additional baseline covariates which are to be adjusted for, but to simplify the key points we do not include these.

For each patient, some measurements may be missing. We note that depending on the chosen estimand, the actual outcomes may be observed, but the potential outcomes of interest under the chosen estimand are missing. We assume that only monotone missingness occurs, although we note that implementations of J2R typically handle intermediate missingness using MAR imputation. Let D_i denote the time of the last observation for patient i. A patient with complete follow-up thus has $D_i = J$. The MAR assumption says that the probability of each pattern of missingness occurring depends only on the data observed under that pattern (Tsiatis 2006), which here means

$$P(D_i = j | Y_{i0}, Y_{i1}, \ldots, Y_{iJ}, X_i) = P(D_i = j | Y_{i0}, Y_{i1}, \ldots, Y_{ij}, X_i),$$

that is, that the (marginal) probability of dropping out immediately after time j does not (statistically) depend on $Y_{j+1}, \ldots, Y_J$, conditional on the randomized treatment X_i and outcome measurements obtained up to and including time j. Alternatively, as described by Daniels and Hogan (2008) it can be equivalently stated as

$$P(D_i = j | D_i \geq j, Y_{i0}, Y_{i1}, \ldots, Y_{iJ}, X_i)$$
$$= P(D_i = j | D_i \geq j, Y_{i0}, Y_{i1}, \ldots, Y_{ij}, X_i),$$

for $j = 1, \ldots, J - 1$. That is, among those who have not yet "dropped out" at time j, the probability that they drop out before time $j+1$ does not depend on the outcomes $Y_{j+1}, \ldots, Y_J$, conditional on treatment group and the outcomes measured through to time j.

Multiple imputation as originally conceived is based on taking a Bayesian approach, with noninformative priors specified for the imputation model parameters. In general, to create a single imputed dataset in this Bayesian paradigm, step one consists

of taking a draw from the observed data posterior distribution of the imputation model parameter. This sometimes requires use of Markov chain Monte Carlo (MCMC) methods. Step two consists of drawing from the conditional distribution of the missing data given the observed, with the model parameter set to the value drawn in step one.

In J2R MI, in the reference arm we assume that $\mathbf{Y}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{iJ})$ is distributed multivariate normal with a distinct mean at each follow-up time and an unstructured covariance matrix, and this model is fitted to the observed data using maximum likelihood, assuming MAR. The same model is separately fitted to the data in the active arm, with distinct parameters. To generate an imputed dataset, from each model a posterior draw is taken of the respective model parameters. Following Carpenter, Roger, and Kenward (2013), we let $\boldsymbol{\mu}_r = (\mu_{r,0}, \ldots, \mu_{r,J})^T$ and unstructured covariance matrix Σ_r denote the resulting posterior draws of the mean and covariance matrix in the reference arm, and $\boldsymbol{\mu}_a$ and Σ_a for the corresponding parameters in the active arm. We then partition the covariance matrices at time D_i as

$$\Sigma_r = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$
$$\Sigma_a = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

Consider a patient with observed values $Y_{i0}, Y_{i1}, \ldots, Y_{iD_i}$ and missing values $Y_{i(D_i+1)}, \ldots, Y_{iJ}$. If they are in the reference arm, then their missing values are imputed from the conditional distribution implied by the assumed model under MAR and the reference arm model parameter posterior draws $\boldsymbol{\mu}_r$ and Σ_r. If the patient is in the active treatment arm, then they are imputed using the multivariate normal conditional distribution implied by assuming that their full data vector has marginal mean equal to

$$\tilde{\boldsymbol{\mu}}_i = (\mu_{a,0}, \ldots, \mu_{a,D_i}, \mu_{r,D_i+1}, \ldots, \mu_{r,J})^T.$$

and variance covariance equal to

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{bmatrix},$$

where

$$\tilde{\Sigma}_{11} = \mathbf{A}_{11}$$
$$\tilde{\Sigma}_{21} = \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{A}_{11}$$
$$\tilde{\Sigma}_{22} = \mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}(\mathbf{R}_{11} - \mathbf{A}_{11})\mathbf{R}_{11}^{-1}\mathbf{R}_{12},$$

The latter values are those that ensure that the sub-matrix of $\tilde{\Sigma}$ corresponding to the observed measurements matches that in the active arm and the conditional covariance matrix of the missing components given the observed matches that in the reference arm.

The preceding describes the J2R approach. Carpenter, Roger, and Kenward (2013) also proposed a number of other variants, including last mean carried forward (LMCF), copy increments in reference and CR.

2.2. Uncongeniality of Reference-Based Imputation

Seaman, White, and Leacy (2014) and Lu (2014) both noted that reference-based imputation approaches were uncongenial with what would be the standard analysis of the resulting imputed data, namely a linear regression model of the final time point outcome, with randomized treatment as covariate (plus other baseline covariates typically). As such, Rubin's variance estimator is not unbiased (even asymptotically) for the true repeated sampling variance of the estimator of treatment effect obtained after using reference-based MI to impute missing values.

To investigate why reference-based MI leads to uncongeniality, we first review the definition of congeniality (Meng 1994; Xie and Meng 2017). We let $\mathbf{Z}_{com} = \{(X_i, D_i, Y_{i0}, \ldots, Y_{iJ}); i = 1, \ldots, n\}$ denote the complete data, $\mathbf{Z}_{mis}$ denote the missing component, and $\mathbf{Z}_{obs}$ denote the observed component. Let θ denote the parameter of interest, which here is the difference in mean outcomes between randomized groups at the final time point, $\theta = E(Y_J|X = 1) - E(Y_J|X = 0)$. In the context considered by Meng (1994), the imputer and analyst in general are distinct entities, whereas in the present context of clinical trials they are the same person. In the absence of any other baseline covariates, the analysis model is a linear regression model for Y_J with Y_0 and X as covariates. In Meng's terminology, the "analyst's complete data procedure" is the ordinary least squares estimator of the coefficient of X in this regression, which we denote $\hat{\theta}^A(\tilde{\mathbf{Z}}_{com})$, where A stands for analyst and $\tilde{\mathbf{Z}}_{com}$ is an arbitrary complete dataset. The analyst's variance estimator, $W^A(\tilde{\mathbf{Z}}_{com})$, is the standard model based variance estimator from linear regression for the coefficient of randomized treatment group X.

Following Xie and Meng (2017) and Bartlett and Hughes (2020), the imputation model and the analyst's complete data procedure are said to be congenial if there exists a unifying Bayesian model (referred to by IA) which embeds the imputer's imputation model and the analyst's complete data procedure, in the sense that

1. For all possible complete datasets $\tilde{\mathbf{Z}}_{com}$,

$$\hat{\theta}^A(\tilde{\mathbf{Z}}_{com}) = E^{IA}(\theta|\tilde{\mathbf{Z}}_{com}) \text{ and}$$
$$W^A(\tilde{\mathbf{Z}}_{com}) = \text{var}^{IA}(\theta|\tilde{\mathbf{Z}}_{com}) \tag{1}$$

where E^{IA} and var^{IA} denote posterior expectation and variance with respect to the embedding Bayesian model;

2. For all possible $\tilde{\mathbf{Z}}_{mis}$,

$$f^I(\tilde{\mathbf{Z}}_{mis}|\mathbf{Z}_{obs}) = f^{IA}(\tilde{\mathbf{Z}}_{mis}|\mathbf{Z}_{obs}) \tag{2}$$

where $f^I(\tilde{\mathbf{Z}}_{mis}|\mathbf{Z}_{obs})$ denotes the predictive distribution for the missing data used by the imputation model and $f^{IA}(\tilde{\mathbf{Z}}_{mis}|\mathbf{Z}_{obs})$ is the predictive distribution for the missing data given the observed data under the embedding Bayesian model.

To see more clearly why reference-based MI leads to uncongeniality, following Seaman, White, and Leacy (2014) and Carpenter et al. (2014), we consider an unrealistic but instructive situation where we omit the baseline measurement Y_0 and set $J = 1$. Thus now $D = 0$ indicates that the single outcome Y is missing and $D = 1$ indicates that Y is observed. With no baseline measurement, the analyst's complete data procedure reduces to calculating the difference in mean of Y between those randomized to active ($X = 1$) and those randomized to reference ($X = 0$):

$$\hat{\theta}^A(\tilde{\mathbf{Z}}_{com}) = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i} - \frac{\sum_{i=1}^n Y_i(1 - X_i)}{\sum_{i=1}^n 1 - X_i} \tag{3}$$

The J2R imputation model in this highly simplified case assumes that

$$Y|D = 0, X = 0 \sim N(\mu_r, \sigma_r^2),$$
$$Y|D = 1, X = 0 \sim N(\mu_r, \sigma_r^2),$$
$$Y|D = 0, X = 1 \sim N(\mu_r, \sigma_r^2),$$
$$Y|D = 1, X = 1 \sim N(\mu_a, \sigma_a^2), \tag{4}$$

such that all outcomes have mean μ_r except those in the active arm with $D = 1$, who have mean μ_a. The J2R imputation model for the missing data can be embedded in a model for the complete data with $f(Y, D, X) = f(Y|D, X)P(D|X)P(X)$ in which $f(Y|D, X)$ is given by the normal models in equation (4) and $P(D = 0|X = x) = \pi_x$, $x = 0, 1$, so that $E(D|X = x) = 1 - \pi_x$. We do not specify $P(X)$ (although we know its distribution from the randomization scheme), but rather perform inference conditional on X. With this embedding model, the J2R imputation procedure satisfies the second part (Equation (2)) of the congeniality definition by construction.

Given an arbitrary complete dataset $\tilde{\mathbf{Z}}_{com}$, under the model embedding J2R imputation the MLE of μ_r is the mean outcome combining the reference arm patients with those in the active arm with $D = 0$, which can be expressed as:

$$\hat{\mu}_r^{com} = \frac{\sum_{i=1}^n Y_i(1 - D_i X_i)}{\sum_{i=1}^n 1 - D_i X_i},$$

and the MLE of μ_a is the mean outcome in those with $X = 1$ and $D = 1$:

$$\hat{\mu}_a = \frac{\sum_{i=1}^n Y_i D_i X_i}{\sum_{i=1}^n D_i X_i}$$

Lastly, the MLE of π_x for $x = 0, 1$ is simply the sample proportion with $D = 0$ in the $X = 0$ and $X = 1$ treatment groups. Then under the embedding model we have that

$$\theta = E(Y|X = 1) - E(Y|X = 0)$$
$$= E[E(Y|X = 1, D)|X = 1] - E[E(Y|X = 0, D)|X = 0]$$
$$= E[\mu_r + (\mu_a - \mu_r)D|X = 1] - E(\mu_r|X = 0)$$
$$= \mu_r + (\mu_a - \mu_r)E(D|X = 1) - \mu_r$$
$$= (\mu_a - \mu_r)(1 - \pi_1). \tag{5}$$

The MLE of θ given complete data under the embedding model follows from this expression under the invariance property of MLE. Morever, for large n the posterior mean under the embedding model is (essentially) equal to the MLE, so that for large n we have

$$E^{IA}(\theta|\tilde{\mathbf{Z}}_{com}) = (\hat{\mu}_a - \hat{\mu}_r^{com})(1 - \hat{\pi}_1)$$
$$= \left(\frac{\sum_{i=1}^n Y_i D_i X_i}{\sum_{i=1}^n D_i X_i} - \frac{\sum_{i=1}^n Y_i(1 - D_i X_i)}{\sum_{i=1}^n 1 - D_i X_i}\right)$$
$$\frac{\sum_{i=1}^n D_i X_i}{\sum_{i=1}^n X_i}, \tag{6}$$

which, unlike the analyst's complete data estimator, uses D, and is not equal to the analyst's complete data estimator $\hat{\theta}^A(\tilde{\mathbf{Z}}_{\text{com}})$. Indeed, suppose that in the active arm virtually all patients were missing their outcomes, such that $\hat{\pi}_1 \approx 1$. In this case $E^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}}) \approx 0$, whereas the analyst's complete data estimator is not (in general). Thus the first part of the first condition in the congeniality definition is not satisfied.

Despite the fact that the analyst's complete data estimator does not in general match the complete data posterior mean under the embedding model, the J2R MI estimator for θ which uses the analyst's complete data estimator *is* equivalent (with an infinite number of imputations) to the Bayesian posterior mean under the embedding model. To see this, following Seaman, White, and Leacy (2014) and Carpenter et al. (2014), consider the MLE/posterior mean of θ given the observed data under the model embedding J2R imputation. The only change moving from the complete data to the observed data is that the MLE of μ_r is now based only on reference arm patients with $D = 1$, so that

$$E^{IA}(\theta|\mathbf{Z}_{\text{obs}}) = (\hat{\mu}_a - \hat{\mu}_r^{\text{obs}})(1 - \hat{\pi}_1). \quad (7)$$

As the number of imputations goes to infinity, and provided n is sufficiently large for the priors to have essentially no impact, under J2R MI the active group mean converges to $(1 - \hat{\pi}_1)\hat{\mu}_a + \hat{\pi}_1\hat{\mu}_r^{\text{obs}}$, whereas the control group mean converges to $\hat{\mu}_r^{\text{obs}}$. Thus the J2R MI estimator of θ converges to (as the number of imputations tends to infinity)

$$(1 - \hat{\pi}_1)\hat{\mu}_a + \hat{\pi}_1\hat{\mu}_r^{\text{obs}} - \hat{\mu}_r^{\text{obs}} = (\hat{\mu}_a - \hat{\mu}_r^{\text{obs}})(1 - \hat{\pi}_1)$$
$$= E^{IA}(\theta|\mathbf{Z}_{\text{obs}}). \quad (8)$$

Turning to the variance, we must check whether the analyst's complete data variance $W^A(\tilde{\mathbf{Z}}_{\text{com}})$ matches the posterior variance given complete data under the embedding model. The analyst's complete data variance estimator could either assume equal variances for Y in the two groups (i.e., the standard t-test) or could use the variance estimator which relaxes this assumption, by estimating the variance separately in each group. Like Carpenter et al. (2014), we will assume the analyst does the latter, so that

$$W^A(\tilde{\mathbf{Z}}_{\text{com}}) = \frac{\widehat{\text{var}}(Y|X = 1)}{n_a} + \frac{\widehat{\text{var}}(Y|X = 0)}{n_r}, \quad (9)$$

where n_a and n_r denote the number randomized to active and control, and $\widehat{\text{var}}(Y|X = 1)$ and $\widehat{\text{var}}(Y|X = 0)$ are the sample variances in each treatment group. Assuming the J2R assumptions (Equation (4)), we have $\text{var}(Y|X = 0) = \sigma_r^2$. For $\text{var}(Y|X = 1)$, we can use the law of total variance to give that

$$\text{var}(Y|X = 1) = \text{var}[E(Y|X = 1, D)|X = 1]$$
$$+ E[\text{var}(Y|X = 1, D)|X = 1]$$
$$= \text{var}[\mu_r + (\mu_a - \mu_r)D|X = 1]$$
$$+ E[\sigma_a^2 D + \sigma_r^2(1 - D)|X = 1]$$
$$= (\mu_a - \mu_r)^2\pi_1(1 - \pi_1) + \sigma_a^2(1 - \pi_1) + \sigma_r^2\pi_1. \quad (10)$$

For the complete data posterior variance under the embedding model, again suppose n is large so that this matches the MLE

estimated variance based on the observed information matrix. Some algebra shows that this is equal to

$$\text{var}^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}}) = (1 - \hat{\pi}_1)$$
$$\left[\frac{\hat{\sigma}_r^2(1 - \hat{\pi}_1)}{n_r + n_a\hat{\pi}_1} + \frac{\hat{\sigma}_a^2}{n_a} + \frac{(\hat{\mu}_r^{\text{com}} - \hat{\mu}_a)^2\hat{\pi}_1}{n_a}\right], \quad (11)$$

where $\hat{\sigma}_a^2$ is the estimated variance of Y from those with $X = 1$ and $D = 1$ and $\hat{\sigma}_r^2$ is the estimated variance from the remaining patients (i.e., $X = 0$, or $X = 1$ and $D = 0$). Equations (9) and (11) are not the same, as required for the second part of the first condition in the definition of congeniality. For example, consider again the case that almost all patients in the active arm have missing data, such that $\hat{\pi}_1 \approx 1$. Then from Equation (6) the complete data posterior mean is approximately zero and from Equation (11) its estimated variance is also approximately zero. In contrast, if $\pi_1 \approx 1$, from equation (10) $\text{Var}(Y|X = 1) \approx \sigma_r^2$, and the analyst's complete data variance estimator of Equation (9) will (on average) estimate $\sigma_r^2(n_a^{-1} + n_r^{-1})$, that is, greater than zero. Thus the second part of the first condition in the congeniality definition is also not satisfied.

Rubin's rules variance estimator is based on decomposing the posterior variance of θ under the embedding model as

$$\text{var}^{IA}(\theta|\mathbf{Z}_{\text{obs}}) = E^{IA}\left[\text{var}^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}})|\mathbf{Z}_{\text{obs}}\right]$$
$$+ \text{var}^{IA}\left[E^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}})|\mathbf{Z}_{\text{obs}}\right]$$

Rubin's rules approximates the first part, the within-imputation variance, by substituting $W^A(\tilde{\mathbf{Z}}_{\text{com}})$ for $\text{var}^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}})$. Since as we have seen $W^A(\tilde{\mathbf{Z}}_{\text{com}})$ is too large, this component will be biased upward. Rubin's rules approximates the second part, the between-imputation variance, by substituting $\hat{\theta}^A(\tilde{\mathbf{Z}}_{\text{com}})$ for $E^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}})$. Consider again the case where $\hat{\pi}_1 \approx 1$. As we have noted previously, from Equation (6), in this case $E^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}}) \approx 0$, and so $\text{var}^{IA}\left[E^{IA}(\theta|\tilde{\mathbf{Z}}_{\text{com}})|\mathbf{Z}_{\text{obs}}\right] \approx 0$. In contrast, the value of $\hat{\theta}^A(\tilde{\mathbf{Z}}_{\text{com}})$ will vary across imputations, so that the estimated between-imputation variance will be larger than zero.

In conclusion, the observed data posterior mean of θ under the embedding model essentially matches the J2R MI estimator of θ. Rubin's rules variance estimator is however larger than the observed data posterior variance under the embedding model. Assuming the embedding model is correct, the latter will (asymptotically) estimate the true frequentist variance of the point estimator correctly, and thus we conclude Rubin's variance estimator is biased upward compared to the true frequentist variance of the point estimator.

Regarding other variants of reference-based MI, positive bias in Rubin's variance estimator for a CR-type approach was shown through simulation by Lu (2014) and Tang (2017) derived analytical expressions for the frequentist bias in Rubin's variance estimator for J2R and CR MI. Gao, Liu, Zeng, Diao, Heyse, and Ibrahim (2017) showed by simulation upward bias in Rubin's variance estimator in the case of repeated binary outcomes. We note that, as shown for example by Robins and Wang (2000), there are other settings where Rubin's variance estimator can be biased downwards compared to the repeated sampling variance of the MI estimator.

3. What is the Right Variance for Reference-Based Multiple Imputation?

As described in the previous section, for reference-based MI, Rubin's variance estimator is biased relative to the true repeated sampling variance of the point estimator of the treatment effect. How large the estimated variance of the treatment effect estimator should be is obviously important, since it affects the advertized precision of the estimated treatment effect and consequently Type I error and power to detect an effect. As expected given the upward (frequentist) bias in Rubin's variance estimator, simulation studies have shown that use of reference-based MI with Rubin's rules leads to conservative Type I error control under the null and the potential for substantial power loss (compared to using the frequentist variance) under the alternative (Gao, Liu, Zeng, Diao, Heyse, and Ibrahim 2017; Lu 2014; Tang 2017).

Carpenter et al. (2014) argued that the frequentist repeated sampling variance is inappropriate in the context of using reference-based MI as a sensitivity analysis to a primary analysis which handles the missing data under a "baseline" assumption, for example, MAR. They proposed a principle that for missing data sensitivity analyses, the variance should be no lower (on average) than the complete data variance estimator, and they showed that reference-based MI with the frequentist variance violates this principle.

Cro, Carpenter, and Kenward (2019) developed this principle further, considering a trial in which missing data are first handled using a "primary" set of assumptions about missingness and second handled using an alternative "sensitivity" set of assumptions. They defined an information anchored sensitivity analysis (e.g., an analysis using J2R) as one in which the relative loss in information about θ caused by missing data is the same as the loss in the primary analysis. Cro, Carpenter, and Kenward (2019) argued that in the context of trials, information anchored sensitivity analyses seem appropriate because, relative to the primary analysis assumptions, they are neither adding nor removing information. They showed that reference-based MI such as J2R are approximately information anchored when Rubin's variance estimator is used for inferences, whereas the repeated sampling variance is information positive—information is being added relative to an MAR-based analysis.

In contrast, others have argued that the repeated sampling variance may be more appropriate. White, Joseph and Best (2020) and Gao, Liu, Zeng, Diao, Heyse, and Ibrahim (2017) pointed out that using reference-based MI with Rubin's rules leads to Type 1 error rates which are too small under the null and a loss of power under the alternative, and as such when used for the primary analysis, the frequentist variance may be preferable.

Consider the use of reference-based MI as a primary analysis estimator of treatment effect. Cro, Carpenter, and Kenward (2019) quite reasonably point out that it seems very counterintuitive to use a method which apparently is able to make more precise inferences the more data is lost. As we have seen however, this is a logical consequence of the strength of the assumption made by reference-based MI methods. If such behavior is viewed as undesirable, and we believe that in many settings it may be, then we believe the correct response is to conclude that the assumptions made by the reference-based approach are inappropriate, rather than to assign blame to a variance estimator. Indeed, the uncongeniality issue here is caused by the fact we are happy to make a (strong) assumption in the imputation model but not in the analysis. If we truly believe in the assumption or at least want to perform an analysis supposing it is true, then we should use it throughout our analysis (i.e., at both imputation and analysis stages). If we do not believe in it, or feel it is too strong, then we should not use it.

Turning next to the use of reference-based MI for sensitivity analyses, we agree with Cro, Carpenter, and Kenward (2019) that ensuring that sensitivity analyses do not inject or take away information (precision) relative to a primary set of missing data assumptions seems like a reasonable principle to adhere to. However, when considering this statement we believe it is critical to be careful about the precise meaning of 'information'. Cro, Carpenter, and Kenward (2019) implicitly take the view that information corresponds to *estimates* of the variance of point estimators, rather than the true repeated sampling variance of the point estimators. Relative to an MAR analysis, reference-based MI estimators such as J2R *do* inject information when information is judged in terms of the true repeated sampling variance of the estimator. Using reference-based MI with Rubin's rules to estimate the variance in our view amounts to pretending reference-based assumptions about missing data are information anchored to an MAR analysis, when in actual fact they are information positive.

In summary, we believe that under a frequentist inference paradigm, information (precision) should be judged in terms of the true repeated sampling variance of estimators. If one wishes to perform information anchored sensitivity analyses, then we believe the correct solution is to construct missing data assumptions which differ to those made by the primary analysis but which genuinely neither add nor remove information, with information being judged in terms of the estimator's true repeated sampling variance. Cro, Carpenter, and Kenward (2019) propose one possible route to this—adding additional random noise to the reference-based MI estimator so that its true repeated sampling variance matches the primary analysis method's variance. A drawback of this approach is that it would then seem difficult to readily communicate the totality of the missing data assumptions made by such a "added noise" reference-based MI estimator. We thus believe further research is warranted to develop sensitivity analysis methods which are information anchored in the sense described above but which like reference-based methods can be relatively easily communicated.

4. Estimating the Repeated Sampling Variance

In this section we review methods for estimating the frequentist variance of reference-based MI estimators, considering their relative advantages and disadvantages.

4.1. Analytical Variance Estimators

A number of authors have developed analytical variance estimators for reference-based estimators for various endpoint types. For a continuous endpoint (Lu 2014) developed a maximum likelihood estimator with delta method variance under a CR

assumption. Tang (2015) derived equivalent matrix versions of Lu's (2014) CR estimator and accompanying delta method variance estimator. Tang (2017) derived analytical variance estimators for J2R and CR methods. Gao, Liu, Zeng, Diao, Heyse, and Ibrahim (2017) applied the general theory developed by Robins and Wang (2000) to derive analytical variance estimators for reference-based MI in the setting of repeated binary data. In all the preceding articles, the analytical variance estimators show good Type 1 error control under the null in simulations, and improved power under the alternative compared to using Rubin's rules.

Analytical variance estimators have the major advantage of being computationally fast. Their drawback however is that they must be derived specifically for each case and implemented in software. Moreover, as noted by Gao, Liu, Zeng, Diao, Heyse, and Ibrahim (2017), there are situations where it may be difficult to derive such variance estimators, for example when intermediate missing values are imputed assuming MAR in a first stage followed by use of reference-based imputation, or perhaps when different imputation strategies are used for different types of intercurrent event.

4.2. *Congenial Bayesian Approach*

An alternative approach is to perform (congenial) Bayesian inference for the treatment effect under a model which embeds the reference-based assumptions. This approach was developed by Lu (2014) and Liu and Pang (2016). Since it is relatively straightforward to obtain posterior draws of the MAR MMRM models using existing software, provided one can express the treatment effect (under the assumed reference-based assumption) as a function of the model parameters (Equation (5) being an example), one can obtain posterior draws of the treatment effect under this assumption by simply applying this function to the posterior draws of the MMRM model parameters. For large n this approach results in accurate frequentist inferences, provided the assumed model is correct. We emphasize that here congeniality is not an issue here because one constructs the expression for the treatment effect under the assumed reference-based assumpion—there is no uncongenial analyst complete data estimator as there is with the reference-based MI approach.

A possible drawback with this approach however is that, like analytical variance estimators, expressing the treatment effect as a function of the MMRM model parameters may become complex and setup specific, for example if one wanted to make a variety of different imputation assumptions to handle different types of intercurrent events.

4.3. *Bootstrap Variance Estimators*

As noted previously, analytical variance estimators and the congenial Bayesian approach require problem specific derivations and implementations. An alternative which avoids this, at the expense of computational cost, is to use bootstrapping. Gao, Liu, Zeng, Diao, Heyse, and Ibrahim (2017) proposed applying nonparametric bootstrapping to reference-based MI estimators in the context of repeated binary endpoints. Simulations showed it gave Type I error control close to the nominal level under the null and superior power to using Rubin's rules under the alternative. Gaot, Liu, Zeng, Xu, Lin, Diao, Golm, Heyse, and Ibrahim (2017) and Diao et al. (2020) similarly used the same bootstrapping approach for reference-based MI estimators for recurrent event data, while Quan et al. (2018) examined its use for reference-based MI with continuous endpoints. Zhang, Golm, and Liu (2020) examined the performance of bootstrapping when used with a return to baseline type MI approach. We emphasize that under uncongeniality it is critical for the bootstrapping to be applied first, followed by multiple imputation. Approaches based on first multiply imputing missing data and then bootstrapping are not generally valid under uncongeniality (Bartlett and Hughes 2020). Also, it is important to note that consistency of bootstrap variance estimators relies on the statistical functional in view being smooth in suitable sense (Shao and Tu 1995). While these conditions have not been verified for reference-based MI estimators, there is extensive empirical evidence supporting that the nonparametric bootstrap variance estimator can deliver accurate frequentist inference when used with reference-based MI methods.

A major practical drawback to using the bootstrap is its large computational cost. Whereas standard MI is often performed using a relatively small number of imputations, bootstrapping requires a much larger number of bootstraps to give accurate inferences. Moreover, in most implementations of MI, draws from the observed data posterior distribution of the model parameters are required in order to make the imputations "proper" and for Rubin's variance estimator to be valid. For certain models this may require running computationally intensive methods such as Markov Chain Monte-Carlo (MCMC), which also necessitate consideration of how many iterations to run the chains for to achieve stationarity and independence of draws. However, if bootstrap is used for inference, this can all be avoided since bootstrapping does not rely on Rubin's rule: it suffices to generate each imputed dataset conditional on efficient estimates (e.g., MLE) of the imputation model parameters, as proposed by von Hippel and Bartlett (2021). To implement this one needs to make a generally minor modification to existing software, by skipping the step in the algorithm that performs the posterior draw.

One might be tempted to reduce computational cost further by reducing the number of imputations performed on each bootstrap sample. This however increases the Monte-Carlo noise in the treatment effect estimator, leading to a somewhat less precise effect estimate and wider confidence intervals than are necessary (Bartlett and Hughes 2020). An alternative bootstrap approach which overcomes this issue was proposed by von Hippel and Bartlett (2021). Their approach performs a small (e.g., two) number of imputations of each bootstrap. The point estimator is taken as the average of estimates across all bootstraps and imputations. To estimate the variance of this estimator, they fit a simple random intercepts model to estimate the between bootstrap and within-bootstrap (between imputation) variances. Bartlett and Hughes (2020) demonstrated through simulation that the approach of von Hippel and Bartlett (2021) provided efficient frequentist valid inferences under uncongeniality yet is substantially quicker to run compared to applying standard nonparametric bootstrapping to an MI estimator which uses a large number of imputations.

5. Simulations

In this section, we report the results of simulations to demonstrate the performance of Rubin's rules and the von Hippel bootstrap approach to reference-based imputation methods. We consider the case of recurrent event data. Trials of one-year duration were simulated with 250 patients in each treatment group. Event counts were simulated using a negative binomial model with the R package dejaVu. Under the null hypothesis event rates were specified as 0.01 (on the days time scale) in both treatment groups, while under the alternative hypothesis the event rate was 0.01 in the control arm and 0.005 in the active arm. A dispersion parameter of 0.25 was used for all scenarios. Dropout completely at random was simulated with each patient's dropout time a draw from an exponential distribution either with rate 0.00025 (leading to 91% of patients completing follow-up) or 0.0025 (leading to 40% of patients completing follow-up).

For those patients who dropped out, the event count in the missing part of their follow-up was imputed either using J2R or CR assumptions as described by Keene et al. (2014). For MI with Rubin's rules ten imputations were generated. Each imputed dataset was generated conditional on a draw from the approximate posterior distribution of the model parameters. For the von Hippel bootstrap approach, two imputations were generated for each bootstrap sample (200 bootstraps), conditional on the maximum likelihood estimates of the fitted model to the corresponding bootstrap sample, as described by von Hippel and Bartlett (2021). In both cases, a negative binomial regression model was fitted to each of the imputed datasets, with treatment as covariate. R code used for the simulations is available at https://github.com/jwb133/refBasedVar.

Table 1 shows the average log rate ratio estimates for each method under the null hypothesis, the empirical standard error (SE) (deviation) of these estimates ("Emp. SE") and the average of the method's estimated SE ('Est. SE'). The final column indicates the ratio of the SEs. Results are based on 5000 simulations. In the low dropout scenario, both Rubin's rules and bootstrap SEs had small bias relative to the empirical SE. However, for the large dropout scenario, Rubin's rules SE was biased upward, for both J2R and CR, compared to the empirical SE. In contrast, the bootstrap SEs remained essentially unbiased. In such a scenario, use of the Rubin's rules SEs would lead to Type 1 being controlled a level lower than the nominal level.

Table 1. Simulation results from 5000 simulations under the null hypothesis.

Method	Est. log RR	Emp. SE	Est. SE	SE ratio
Low dropout				
Rubin J2R	0.000	0.044	0.047	1.058
Bootstrap J2R	0.000	0.044	0.045	1.012
Rubin CR	0.000	0.045	0.047	1.047
Bootstrap CR	0.000	0.045	0.045	1.011
High dropout				
Rubin J2R	−0.001	0.036	0.052	1.454
Bootstrap J2R	0.000	0.035	0.035	1.005
Rubin CR	−0.001	0.039	0.052	1.331
Bootstrap CR	−0.001	0.039	0.039	1.005

NOTE: Mean estimated log rate ratio (Est. log RR), empirical standard error (Emp. SE), mean estimated standard error (Est. SE) and ratio of mean estimated SE to empirical SE (SE ratio).

Table 2. Simulation results from 5000 simulations under the alternative hypothesis.

Method	Est. log RR	Emp. SE	Est. SE	SE ratio
Low dropout				
Rubin J2R	−0.650	0.049	0.053	1.089
Bootstrap J2R	−0.650	0.049	0.049	0.998
Rubin CR	−0.659	0.050	0.053	1.068
Bootstrap CR	−0.659	0.050	0.050	0.997
High dropout				
Rubin J2R	−0.398	0.034	0.060	1.754
Bootstrap J2R	−0.398	0.033	0.032	0.982
Rubin CR	−0.447	0.039	0.060	1.516
Bootstrap CR	−0.446	0.038	0.038	0.982

NOTE: Mean estimated log rate ratio (Est. log RR), empirical standard error (Emp. SE), mean estimated standard error (Est. SE) and ratio of mean estimated SE to empirical SE (SE ratio).

Table 2 shows the results under the alternative scenario. In the low dropout scenario, the J2R and CR estimated effects were close to the effect that would be seen under MAR (i.e., a log rate ratio of $\log(0.5) = -0.69$). Under the high dropout scenario, the treatment effect was diluted considerably toward the null, with CR less conservative than J2R, as has been seen previously by others. Again both Rubin's rules and bootstrap gave approximately unbiased SEs in the low dropout scenario, while again Rubin's rules SEs were biased upward substantially and the bootstrap SEs were unbiased. In this scenario use of the bootstrap SEs would lead to materially increased power to detect a treatment effect.

6. Discussion

It has been known for almost 10 years that Rubin's variance estimator is biased upward relative to the true repeated sampling variance of reference-based estimators of treatment effects in randomized trials, but there remains no settled view on what is the right variance to use. Given the increasing use of reference-based MI in trials, this is not particularly satisfactory. We have argued that the frequentist variance is the correct variance for reference-based estimators. If the behavior of the frequentist variance does not seem appropriate to the analyst, for example, because it decreases as the amount of missing data in the active arm increases, then our view is that this means the analyst does not really belief the assumptions made by the reference-based approach.

In the context of performing missing data sensitivity analyses, the proposed principle that sensitivity analyses should be information neutral, or anchored, seems eminently sensible. However, we believe that provided we are operating under the frequentist paradigm, information here must be judged in terms of repeated sampling variance. In our view, using reference-based MI with Rubin's variance estimator amounts to using a point estimator that is not information anchored but then using a variance estimator that falsely suggests it is information anchored. Further research is warranted to develop new sensitivity analyses which retain the attractive features of a reference-based type approach, where assumptions are structured qualitatively, rather than quantitatively, but which are information anchored in the frequentist variance sense.

Historically, reference-based estimators have tended to be mostly used as sensitivity analyses to a primary analysis which

adopts the MAR assumption. However, such methods, and combinations of them are being increasingly used to develop estimators which might be used as the primary analysis method of trials (e.g., Darken et al. 2020). In this context, it is clearly important to assess whether the strong assumptions potentially made by the missing data assumptions of such approaches are justifiable, particularly in light of what they imply for the repeated sampling variance of the treatment effect estimator.

We have suggested that a particular way of combining bootstrapping with MI can be used to obtain estimates of frequentist variance of reference-based MI estimators. Because Rubin's rules are no longer used, the imputation process does not need to be "proper," and imputation can instead be performed conditional on maximum likelihood estimates. As described by von Hippel and Bartlett (2021), to implement this in software should in most cases require relatively small changes, since the part that performs the draw from the posterior distribution (e.g., via MCMC sampling) can simply be skipped. In R, J2R MI is implemented for continuous endpoints using this approach in the `mlmi` package, while the `RefBasedMI` package (*https://github.com/UCL/RefbasedMI*), currently under development, has an option that allows the user to impute conditional on the MLE for a much wider range of reference-based assumptions for continuous endpoints. The R package `dejaVu` implements reference-based MI for recurrent event data, as proposed by Keene et al. (2014), and also includes an option to impute conditional on the maximum likelihood estimates. The calculations to implement the bootstrap/MI approach proposed by von Hippel and Bartlett (2021) are relatively simple, but for R users they are implemented in the `bootImpute` package.

Acknowledgments

Funding

ORCID

Jonathan W. Bartlett ⬤ https://orcid.org/0000-0001-7117-0195

References

Atkinson, A., Kenward, M. G., Clayton, T. and Carpenter, J. R. (2019), "Reference-Based Sensitivity Analysis for Time-to-Event Data," *Pharmaceutical Statistics*, 18, 645–658. [2]

Bartlett, J. W., and Hughes, R. A. (2020), "Bootstrap Inference for Multiple Imputation Under Uncongeniality and Misspecification," *Statistical Methods in Medical Research*, 29, 3533–3546. [3,6]

Carpenter, J. R., Roger, J. H., and Kenward, M. G. (2013), "Analysis of Longitudinal Trials With Protocol Deviation: A Framework for Relevant, Accessible Assumptions, and Inference Via Multiple Imputation," *Journal of Biopharmaceutical Statistics* 23, 1352–1371. [1,2]

Carpenter, J., Roger, J., Cro, S., and Kenward, M. (2014), "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,' *Journal of Biopharmaceutical Statistics*, 23, 1352–1371," *Journal of Biopharmaceutical Statistics*, 24, 1363–1369. [1,3,4,5]

Cro, S., Carpenter, J. R., and Kenward, M. G. (2019), "Information-Anchored Sensitivity Analysis: Theory and Application," *Journal of the Royal Statistical Society*, Series A, 182, 623–645. [1,5]

Daniels, M. J., and Hogan, J. W. (2008), *Missing Data in Longitudinal Studies*, Boca Raton, FL: Chapman & Hall / CRC. [2]

Darken, P., Nyberg, J., Ballal, S., and Wright, D. (2020), "The Attributable Estimand: A New Approach to Account for Intercurrent Events," *Pharmaceutical Statistics*, 19 . [8]

Diao, G., Liu, G. F., Zeng, D., Zhang, Y., Golm, G., Heyse, J. F., and Ibrahim, J. G. (2020), "Efficient Multiple Imputation for Sensitivity Analysis of Recurrent Events Data With Informative Censoring," *Statistics in Biopharmaceutical Research*, 1–9, DOI: 10.1080/19466315.2020.1819403. [6]

Gao, F., Liu, G. F., Zeng, D., Xu, L., Lin, B., Diao, G., Golm, G., Heyse, J. F., and Ibrahim, J. G. (2017), "Control-Based Imputation for Sensitivity Analyses in Informative Censoring for Recurrent Event Data," *Pharmaceutical Statistics*, 16, 424–432. [6]

Gao, F., Liu, G., Zeng, D., Diao, G., Heyse, J. F., and Ibrahim, J. G. (2017), "On Inference of Control-Based Imputation for Analysis of Repeated Binary Outcomes With Missing Data," *Journal of Biopharmaceutical Statistics*, 27, 358–372. [1,4,5,6]

Keene, O. N., Roger, J. H., Hartley, B. F., and Kenward, M. G. (2014), "Missing Data Sensitivity Analysis for Recurrent Event Data Using Controlled Imputation," *Pharmaceutical Statistics*, 13, 258–264. [2,7,8]

Liu, G. F., and Pang, L. (2016), "On Analysis of Longitudinal Clinical Trials With Missing Data Using Reference-Based Imputation," *Journal of Biopharmaceutical Statistics*, 26, 924–936. [6]

Lu, K. (2014), "An Analytic Method for the Placebo-Based Pattern-Mixture Model," *Statistics in Medicine*, 33, 1134–1145. [1,3,4,5,6]

Mallinckrodt, C., Bell, J., Liu, G., Ratitch, B., O'Kelly, M., Lipkovich, I., Singh, P., Xu, L., and Molenberghs, G. (2019), "Aligning Estimators With Estimands in Clinical Trials: Putting the ICH E9 (R1) Guidelines Into Practice," *Therapeutic Innovation & Regulatory Science*, 54,353–364. [1]

Meng, X. L. (1994), "Multiple-Imputation Inferences With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 10, 538–573. [1,3]

Quan, H., Qi, L., Luo, X., and Darchy, L. (2018), "Considerations of Multiple Imputation Approaches for Handling Missing Data in Clinical Trials," *Contemporary Clinical Trials*, 70, 62–71. [6]

Robins, J. M., and Wang, N. (2000), "Inference for Imputation Estimators," *Biometrika*, 85, 113–124. [4,6]

Seaman, S. R., White, I. R., and Leacy, F. P. (2014), "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," *Journal of Biopharmaceutical Statistics*, 24, 1358–1362. [1,3,4]

Shao, J., and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer Science & Business Media. New York: Springer. [6]

Tan, P.-T., Cro, S., Van Vogt, E., Szigeti, M. and Cornelius, V. R. (2021), "A Review of the Use of Controlled Multiple Imputation in Randomised Controlled Trials With Missing Outcome Data," *BMC Medical Research Methodology*, 21, 1–17. [1]

Tang, Y. (2015), "Short Notes on Maximum Likelihood Inference for Control-Based Pattern-Mixture Models," *Pharmaceutical Statistics*, 14, 395–399. [6]

——— (2017), "On the Multiple Imputation Variance Estimator for Control-Based and Delta-Adjusted Pattern Mixture Models," *Biometrics*, 73, 1379–1387. [1,4,5,6]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [2]

von Hippel, P. T., and Bartlett, J. W. (2021), "Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws," *Statistical Science*, 36, 400–420. [6,7,8]

White, I., Joseph, R., and Best, N. (2020), "A Causal Modelling Framework for Reference-Based Imputation and Tipping Point Analysis in Clinical Trials With Quantitative Outcome," *Journal of Biopharmaceutical Statistics*, 30, 334–350. [1,5]

Xie, X., and Meng, X.-L. (2017), "Dissecting Multiple Imputation From a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?" *Statistica Sinica*, 27, 1485–1545. [3]

Zhang, Y., Golm, G., and Liu, G. (2020), "A Likelihood-Based Approach for the Analysis of Longitudinal Clinical Trials With Return-to-Baseline Imputation," *Statistics in Biosciences*, 12, 23–36. [6]