Is Using Multiple Imputation (MI) Better Than Complete Case (CC) Analysis For Estimating a Prevalence (Risk) Difference In Randomized Controlled Trials When Binary Outcome Observations Are Missing?

Mavuto Mukaka*[1,2,3,4], Sarah A White[2,3], Dianne J Terlouw[1,3], Victor Mwapasa[2], Linda Kalilani-Phiri[2] and Brian Faragher[3].

[1]Malawi-Liverpool-Wellcome Trust Clinical Research Programme, College of Medicine, University of Malawi, Box 30096, Blantyre 3, Malawi;

[2]Department of Public Health, College of Medicine, University of Malawi, P/Bag 360, Blantyre 3, Malawi;

[3]Liverpool School of Tropical Medicine, Pembroke Place, L3 5QA, Liverpool, UK.

[4]Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University,60th Anniversary Chalermprakiat Building, 3rd Floor, 420/6 Ratchawithi Rd, Bangkok,10400, Thailand

Author email addresses; Sarah White (sarah.e.m.white@gmail.com), Dianne Terlouw (anja.terlouw@lstmed.ac.uk), Victor Mwapasa (vmwapasa@gmail.com), Linda Kalilani-Phiri (lkalilani@hotmail.com), and Brian Faragher (brian.faragher@lstmed.ac.uk)

*Correspondence to Dr. Mavuto Mukaka, Head of Statistics, Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, 60th Anniversary Chalermprakiat Building, 3rd Floor, 420/6 Ratchawithi Rd, Ratchathewi District, Bangkok, 10400, Thailand.
E-mail: mmukaka@gmail.com; mavuto@tropmedres.ac

ABSTRACT

Background
Missing outcomes can seriously impair the ability to make correct inferences from randomized controlled trials (RCTs). Complete case (CC) analysis is commonly used, but reduces sample size and is perceived to lead to reduced statistical efficiency of estimates while increasing the potential for bias. As multiple imputation (MI) methods preserve sample size, they are generally viewed as the preferred analytical approach.

Methods
We examined this assumption, comparing the performance of CC and MI methods to determine risk difference (RD) estimates in the presence of missing binary outcomes. We conducted simulation studies of 5000 simulated datasets with 50 imputations of RCTs with one primary follow-up endpoint at different underlying levels of RD (3%-25%) and missing outcomes (5-30%).

Results
For Missing At Random (MAR) or Completely At Random (MCAR) outcomes, CC method estimates generally remained unbiased and achieved precision similar to or better than MI methods, and high statistical coverage. Missing Not At random (MNAR) scenarios yielded invalid inferences with both methods. Effect size estimate bias was reduced in MI methods by always including group membership even if this was unrelated to missingness. Surprisingly**,** under MAR and MCAR conditions in the assessed scenarios, MI offered no statistical advantage over CC methods.

Conclusion
While MI must inherently accompany CC methods for intention-to-treat analyses, these findings endorse CC methods for per protocol risk difference analyses in these conditions. These findings provide an argument for the use of the CC approach to always complement MI analyses, with the usual caveat that the validity of the mechanism for missingness be thoroughly discussed. More importantly, researchers should strive to collect as much data as possible.

Key words: missing binary outcome, risk difference, complete case analysis, multiple imputation, , missing completely at random, missing at random, missing not at random.

BACKGROUND

The Randomized Controlled Trial (RCT) is considered the gold standard study design for evaluating the efficacy of a treatment or intervention in clinical and epidemiological research [1]. A well-designed and conducted RCT provides an efficient and unbiased estimate of effect size when all observations required by the study protocol have been obtained [1, 2] but difficulties can arise if some observations are missing. Missing outcome observations can create a specific and often considerable challenge for the statistical analysis. Incorrectly handled, these can result in biased and inefficient estimates of effect size, threatening the intrinsic strength of the RCT design and compromising the ability to draw valid inferences from the study findings [3].

Missing observations are least likely to occur at the baseline assessment as many of the observations collected at this time are required not merely to provide a reference against which to measure efficacy but also to ensure that recruited participants meet the RCT inclusion / exclusion criteria. Missing observations tend to occur more frequently at follow-up assessments, when it is not uncommon for the primary outcome measure to be missing for some participants [4]. A considerable number of statistical methods have been, and continue to be, proposed for handling missing observations but as yet universally accepted robust methods for handling missing data in RCTs do not exist [5].

Widely used analysis strategies include methods based on multiple imputation (MI), inverse probability weighting (IPW), doubly robust inverse probability weighting (DR-IPW) and Maximum Likelihood Estimation (MLE). Despite the considerable body of literature on such methods, many researchers continue to use the simplest and most expedient approach of simply excluding from the statistical analyses all participants for whom the outcome measure is missing. This analytical method, commonly referred to as complete case (CC) analysis, is the default approach in many statistical packages [2, 6, 7] . There is, however, well documented evidence that a CC analysis may yield biased and inefficient estimates of effect size especially when the missing data levels are high, irrespective of the type and/or pattern of missingness [2, 8, 9] .

The exact impact of a CC analysis on effect size estimates in different situations is poorly understood and has not been explored in detail [10]. This is a key gap in our knowledge. The choice of the most appropriate analytical method for handling missing outcome data in any RCT is ideally informed by the mathematical properties of the different analysis methods available, the missing observation pattern present, and an understanding of the mechanism(s) that led to the missing observations [11]. Furthermore, it cannot be assumed that increased methodological complexity leads to less bias; there are known situations in which MI methods produce identical bias levels to a CC analysis [7, 12].

Many RCTs use a binary outcome measure (survived/died, outcome absent/present, treatment failure/success), in which case effect size is estimated using an odds ratio (OR), risk ratio (RR) or risk difference (RD) [13] . The risk difference is becoming increasingly popular due to its ease of interpretation. Several simulation studies have compared methods for handling missing binary outcome observations when effect size is estimated using an OR [2, 12, 14] but we are not aware of any publications on how missing observation methods perform when effect size is estimated using a RD. As OR and RD modelling use different mathematical algorithms, the results from an OR model cannot necessarily be extrapolated to a RD model.

In this paper, we use simulation methods to compare the performance of CC and MI to estimate effect size using the risk difference in RCTs with missing binary outcome observations and explore which method is preferable for various missing observation patterns and effect size levels.

METHODS

Simulated data sets were generated to compare the impact of CC and MI analytical approaches on effect size estimation in a two-group RCT with a binary outcome measure when some outcome observations are missing, across a range of effect sizes and missing outcome levels as detailed below. The parameters used in each simulated data set were based on the results of a malaria efficacy RCT conducted in Malawi between 2003 and 2006 [15].

For each effect size and missing outcome level combination examined, 5,000 data sets were simulated. To reflect the parent malaria RCT, the sample size for each dataset was 200 participants, with 100 subjects randomized to the intervention and the control groups respectively. The binary outcome was generated using a logit model to achieve a range of effect sizes (treatment group differences); a random process was then used to delete a pre-specified proportion of outcomes.

For each participant in both treatment groups, baseline values were generated to represent their age, weight (wt), hemoglobin (hb) level and malaria parasitaemia count (para) using a multivariate Normal distribution. Haemoglobin level was generated untransformed, but as body weight, age and parasitaemia counts had skewed distributions in the parent RCT, these were generated using a logarithmic scale, with their parameter values (means, variances and covariance) estimated from the log-transformed variable values in the parent RCT [16]. These variables are generally expected to be related to the outcome (adequate clinical and parasitological response). The matrices of parameters used to simulate the baseline covariate observations were:

$$\mathbf{X}=\begin{bmatrix}\log_e(Age)\\ hb\\ \log(Weight)\\ \log(Parasitaemia)\end{bmatrix},\ \boldsymbol{\mu}=\begin{pmatrix}3.15\\ 9.32\\ 2.40\\ 10.7\end{pmatrix},\ \boldsymbol{\sigma}=\begin{pmatrix}0.42\\ 1.66\\ 0.18\\ 1.50\end{pmatrix}\ \text{and}\ \boldsymbol{\rho}=\begin{pmatrix}1.00 & 0.09 & 0.16 & 0.02\\ 0.09 & 1.00 & 0.4 & 0.2\\ 0.16 & 0.4 & 1.00 & 0.05\\ 0.02 & 0.2 & 0.05 & 1.00\end{pmatrix}$$

where:

$\boldsymbol{X}$ is a vector of the four covariates log(age), haemoglobin (hb), log(body weight) and log(parasitaemia);

$\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are vectors of the mean and standard deviation values respectively for each of log(age), haemoglobin, log(body weight) and log(parasitaemia);

$\boldsymbol{\rho}$ is a matrix of the correlations between each pair combination of the baseline covariates.

To maintain the skewness of these covariates found in the parent RCT, the log-Normal generated variables were transformed (exponentiated) back into their original prior to analysis. The model estimated the (binary) outcome as a function of treatment group, age and haemoglobin.

The binary outcome was then simulated for each of the two groups to achieve the desired efficacy (treatment success) rates using a Bernoulli (π_i) distribution, where π_i is the mean proportion of subjects with treatment success (efficacy) in group i, for i=A, B. This resulted in simulated binary outcome data with π_i success rate (efficacy) in group *i*.

The efficacies of treatments A and B respectively were generated using Bernoulli distributions as follows (Y=1 denotes treatment success and T=treatment):

for response rates of 85% in treatment A vs 60% in treatment B

Y = Bernoulli[Pr(Y=1 | T=A)=0.85]
= Bernoulli[Pr(Y=1 | T=B)=0.60]

for response rates of 98% in treatment A vs 95% in treatment B

Y = Bernoulli[Pr(Y=1 | T=A)=0.98]
= Bernoulli[Pr(Y=1 | T=B)=0.95]

Four different imputation models were considered:

model 1: log(weight), haemoglobin, log(age) and log(parasitaemia) were used to simulate the missing outcome observations;

model 2: log(weight) was excluded leaving just haemoglobin, log(age) and log(parasitaemia);

model 3: group membership was added to the covariates used in model 2;

model 4: all of log(age), haemoglobin, log(age), log(parasitaemia) and group membership were used.

Imputations were conducted using the chained equations procedure [17]. Both 10 and 50 imputations were used in each multiple imputation procedure to provide information on the potential impact of increasing imputation rate.

All simulations and statistical summaries were performed using the Stata for windows software (version SE/11; 4905; Stata Corp; College Station, Texas 77845 USA).

Choice of effect size settings

Two different effect size settings were simulated.

1. *85% for the treatment of interest (group A) and 60% for the control treatment (group B).*

   This scenario is not as unrealistic as it might appear. Relatively large effect sizes of this or a greater magnitude are not uncommon in malaria RCTs (e.g. Bell et al (2008) [15]). Furthermore, resistance is often underestimated when designing such trials so sample size calculations are based on smaller differences than are actually observed. Consequently, sample sizes can be over-estimated, producing statistically significant findings even if the (binary) outcome is missing for as much as 30% of participants. This setting was selected primarily, however, to avoid the model convergence problems that can occur when either group returns an effect rate close to the boundary (either 0% or 100%).

2. *98% for the treatment of interest (group A) and 95% for the control treatment (group B).*

For this second setting, both effect rates were deliberately set close to the boundary value of 100% as this is a common situation in malaria treatment trials comparing highly efficacious Artemisinin-based combination therapies.

Choice of missing outcome settings

Consider an RCT with two treatment arms in which the primary outcome Y is a binary variable measured once, at the end of a fixed period of time of follow-up, for each patient. Let X denote the complete (uni- or multi-dimensional) covariate matrix, and let D be an indicator variable such that D = 1 if Y is missing and D = 0 if Y is observed.

Within this context, the following three missing data mechanisms defined by Rubin [18] were considered.

Outcome Missing At Random (MAR)

An outcome observation was defined as MAR if the probability (Pr) of it being missing was dependent on the observed covariates X but independent of the specific value that theoretically should have been observed for that missing observation [18]. This is expressed mathematically as follows:

$Pr(D = 1 \mid Y, X) = Pr(D = 1 \mid X)$

Outcome Missing Completely At Random (MCAR)

An outcome observation was defined as MCAR if the probability of it being missing was independent of both the observed covariates X and the specific value that theoretically should have been observed for that missing observation [18]. This is expressed mathematically as follows:

$Pr(D = 1 \mid Y, X) = Pr(D = 1)$

Outcome Missing Not At Random (MNAR)

An outcome observation was defined as MNAR if the probability of it being missing was dependent on the observed covariates X, the observed outcome values and the unobserved outcome values [18]. This is expressed mathematically as follows:

$Pr(D = 1 \mid Y, X) = Pr(D = 1 \mid Y^{obs}, Y^{mis}, X)$

where Y^{obs} and Y^{mis} are the observed and missing outcome values respectively.

Method used to simulate MAR, MCAR and MNAR scenarios

Three missing level settings were considered for each scenario: 5%, 15% and 30%.

To generate binary outcome data that were MCAR, a random number with a uniform [0,1] distribution was generated for each participant in the simulated data set. The p% of participants with the smallest random numbers were then coded as having their outcome observation missing, p taking the values 5%, 15% or 30% as appropriate.

The following logistic regression models used to generate missing outcomes with MAR levels of 5%, 15% and 30% respectively and which were dependent on group and weight:

$\text{logit}(\pi) = (0.872 * \text{treatment}) + (0.099 * \text{weight}) - 4.666$

$\text{logit}(\pi) = (0.299 * \text{treatment}) + (0.043 * \text{weight}) - 2.409$

$\text{logit}(\pi) = (0.148 * \text{treatment}) + (0.022 * \text{weight}) - 1.18$

where π is the probability of an outcome being missing.

The models used to generate missing outcomes with MNAR levels of 5%, 15% and 30% respectively were:

$\text{logit}(\pi) = 2.99 * \text{outcome}$; $\text{logit}(\pi) = 1.89 * \text{outcome}$; $\text{logit}(\pi) = 1.20 * \text{outcome}$

The MAR and MNAR missing outcome indicators were thus generated with distributions:

Bernoulli [1 / (1 + exp[1])] for MAR

Bernoulli [1 / (1 + exp{-(b_3 * outcome)})] for MNAR

where b_1, b_2 and b_3 are (regression) coefficients outlined in the missing outcome data logit models above and π is the probability of an outcome being missing.

For MNAR, the models resulted in participants with a successful (positive) outcome being more likely to have their outcome missing, creating a greater proportion of missing outcomes in the high efficacy group than the group with low efficacy, in turn resulting in differential proportions of missing outcomes between the two study groups. This is realistic in the context of malaria trials as successfully treated participants may have less incentive to return for their final assessment, particularly if doing so would be costly or time-consuming.

To minimize the problems of model non-convergence when generated effect sizes are close to the boundary, Cheung's modified least squares method [19] was used to estimate risk differences. This method, which uses ordinary least squares (OLS) estimation together with Huber-White (H-W) robust standard errors, is suitable if interest is confined to the estimation of risk differences but is not suitable theoretically if there is interest in predicting probabilities for individual patients as estimated values can fall outside the probability range 0 to 1.

Data model and model assessment criteria

The outcome of interest was modeled as a function of age, hb and group using the following logistic regression model:

$\text{logit}[P(Y = 1| T=t, hb, age)] = b_0 + b_1*T + b_2*hb + b_3*age,$

where: t = A or B

b_0 , b_1 , b_2, b_3 are estimates of intercept, treatment effect, hb effect and age effect respectively.

Data model specification was identical for all scenarios; only the methods of handling cases that had missing outcomes were varied. For complete case analysis, all cases with missing outcomes were exclude from the statistical analyses.

The performance of the data models from the different approaches of missing data at each level of missing data were compared against three criteria: bias, statistical coverage and Root-Mean-Squared Error (RMSE).

RESULTS

Consistent with the findings of Schafer's and Rubin [20, 21], the results obtained using 10 and 50 imputations were virtually identical, so only the results using 50 imputations are presented.

Missing outcome observations MAR, MCAR and MNAR: 60% vs 85% efficacy (Figure 1)

When the missing outcome setting was MAR, effect size estimates became increasingly inefficient as the proportion of missing outcome observations increased. The RMSE values observed indicated that inefficiency levels were identical for both CC and MI methods.

Using CC methods, effect size estimates were unbiased for all assessed missing value levels. Using MI methods, when group membership was included in the imputation model, only a small degree of bias was observed in the effect size estimates. When group membership was excluded from the model, however, estimates were consistently negatively biased (i.e. effect size was consistently under-estimated), and the degree of bias increased as the proportion of missing outcome values increased.

Coverage was generally high for all models when missing value levels were small or moderate, but when the proportion of missing outcomes was extended to 30%, coverage for the mis-specified MI models fell to around 88% (detailed in appendices 1a-1c). In this context, imputation models not containing both of the variables weight and group membership were technically mis-specified as it was these two variables that determined missingness. MI models containing both weight and group performed well for all missing outcome configurations, providing estimates that were only fractionally biased with good coverage of around 95%, as did MI models that included group but excluded weight. MI models that included weight but excluded group, however, performed as badly as those MI models that included neither weight nor group.

With MCAR, the pattern of results were very similar to those with MAR. Coverage was generally high, remaining close to 95% for all models at all missing value levels, except when the proportion of missing outcomes reached 30%. At this point, as for MAR, coverage fell to around 88% for mis-specified MI models and increased levels of (negative) bias were observed.

Under the MNAR condition, RD estimates contained some degree of (usually but not exclusively positive) bias with both CC and MI methods, the degree of bias rising as the proportion of missing outcome observations increased. Coverage levels tended to be good, but with some deterioration at high missingness levels.

Detailed results for this scenario are provided in appendices 1a, 1b and 1c for MAR, MCAR and MNAR respectively.

Missing outcome observations MAR, MCAR and MNAR: 95% vs 98% efficacy (Figure 2).

When both efficacy levels were close to the 100% boundary, coverage was poorest when there was no missing data (0.939 compared to the set nominal level of 0.950). All complete case

(CC) analyses converged (appendix 2a) but a small proportion of imputed analyses failed to converge or produce output in Stata. Non-convergence occurred more frequently with increasing proportions of missing outcome values (appendix 2a).

With these efficacy levels, all CC analyses converged while a small number of MI analyses failed to converge for all three missing data mechanisms (MAR, MCAR and MNAR). Non-convergence in MI analyses occurred more frequently with increasing proportions of missing outcome values. As the proportion of missing data increased, the standard errors of the effect size estimates increased and the efficiency of all analyses decreased in both CC and MI analyses, though this was less marked with the CC analyses.

The estimates of effect size were unbiased for all missing value levels using CC methods, and only small levels of bias were detected for those imputation models that included group membership for both MAR and MCAR missing data scenarios. For the MAR and MCAR missing data mechanisms, the bias in the effect size estimates was markedly greater using imputation models that did not incorporate group membership.

In the MNAR missing data mechanisms, for the 5% missing level, all the models were generally unbiased with generally good coverage of around 95% - but as missing levels increased all models led to invalid inference of treatment effect. The level of bias increased with increasing missing levels for the CC models and the MI models, even for those that included group membership. The mis-specified models had unbiased estimates but provided invalid inference because the coverage was conservatively too high (close to 100%) instead of the nominal 95%.

Detailed results for this scenario are provided in appendices 2a, 2b and 2c for MAR, MCAR and MNAR respectively.

DISCUSSION

When binary outcome observations are missing that can be assumed to be MAR or MCAR, CC analysis methods were found to perform as well as, and often better than, MI methods, consistently producing unbiased RD estimates. This finding is consistent with those reported by Groenwold [14] who examined missing binary outcomes in a RCT setting using odds ratio as the effect size estimate of interest. The efficient estimates obtained from CC than MI is a counter-intuitive and surprising finding.

The loss of statistical efficiency in CC analyses is be attributed directly to the reduction in the effective sample size that occurs with this method, as has been reported previously [8]. A surprising and unexpected finding, however, was that the loss in efficiency using CC analysis methods was consistently found to be no worse than, and often better than, that observed using MI methods. Theoretically MI methods are expected to yield unbiased standard errors, because sample size is maintained and the uncertainty in the imputed values is fully accounted for [8].

A plausible explanation for this unexpected efficiency finding is that the MI procedures also increase the variability in the outcome values that inflates the standard error of the effect size estimate. This increase in variability is likely caused by the random component that is added to missing outcome values during the imputation process.

No convergence problems were experienced using CC analyses when missing binary outcomes could be assumed to be MAR or MCAR, although some problems were experienced when missingness was MNAR. In contrast, convergence problems occurred under both the MAR and MCAR conditions when imputation models were used, particularly when both efficacy rates were close to the parameter boundaries. This was caused by all imputed values being occasionally allocated to the same outcome value across all imputations when efficacy levels in both groups are close to the boundary, which results in zero standard errors for the effect size estimate, a phenomenon referred to as "perfect prediction" [17]. Perfect prediction can arise in any Generalized Linear Model that has a categorical outcome [22]. The usual reason for perfect prediction in a MI analysis is that all imputed values take the same value for all participants across the imputations, resulting in zero between imputation variance. In this situation the calculation of degrees of freedom would involve division by zero which may result in non-convergence. White has suggested that perfect prediction problems can be a result of the flat likelihood [22].

This problem of "perfect prediction" can be drastically reduced in Stata by using the command option "augment", which causes an augmented regression to be performed [22].

Another striking finding under the MAR and MCAR conditions was that the inclusion of treatment group membership in the imputation process played a crucial role in improving its performance. Excluding this variable from the imputation process produced biased estimates of the adjusted efficacy risk difference. If missingness is related to some covariates, the absence of those covariates in the imputation model appeared to have little impact on bias levels for the effect size estimate, provided group was included in the imputation calculations. Less predictably, including treatment group membership would also appear to be paramount over all other factors even when this is not related to missingness.

Under the MNAR condition, when missingness is related to treatment group membership and outcome, both CC and MI analyses produce biased estimates of effect size. Furthermore, the inclusion of group in a multiple imputation analysis will tend to lead to positive bias away from the null hypothesis. Thus, MNAR binary outcomes appear generally to over-estimate effect size, with the possible exception of mis-specified MI models, a counter-intuitive finding that requires further research. The results from the two mis-specified imputation models are presented in this paper to emphasize that, when assuming MI, care must be taken when selecting the imputation model as using a poor imputation model can bias the effect size estimates.

This study has demonstrated that in the presence of missing binary outcome observations in a RCT with a single follow-up endpoint of interest, CC and MI analysis methods performed very similarly under the three missingness assumptions examined, except when an inappropriate imputation model was adopted, in which case the MI risk difference estimates obtained were generally inferior to those generated by a CC analysis. These findings indicate that, MI methods offered no advantages over the much easier to apply CC method in the scenarios considered.

There are, however, other factors to be considered when analysing the findings of a RCT. The intention to treat principle (ITT) is now the standard procedure for the primary evaluation of a RCT. Under this principle, the use of MI methods may be preferable on the grounds that these retain all patients in the statistical evaluation whereas the CC methods excludes all patients for whom the outcome measure could not be recorded.

A reasonable compromise might be to perform a MI analysis as the primary ITT analysis, following a rigorous exploration of the likely underlying reasons for the missingness in the outcome measure. Inappropriate imputation models can lead to risk difference estimates that are inferior to those from a CC analysis, so a "non-parsimonious" approach to this MI analysis is essential (i.e. as many covariates as possible must be included in the imputation process). In addition, group membership must be included in the imputation model otherwise there is an increased risk of bias even when missingness is not in fact related to group membership. A secondary CC analysis could then be performed as part of the "per protocol" analyses.

MI methods have no place in a "per protocol" risk difference analysis. CC methods yield unbiased effect size estimates and are less prone to the problem of perfect prediction when effect sizes stray close to a boundary. MI methods are more suitable when the missingness is MNAR and thus have an important role both in sensitivity analyses and when the outcome of interest is collected at several points during a study. Unfortunately, the importance of sensitivity analyses is frequently under-valued [23].

From a reporting perspective, the general research community is likely to be skeptical of a statistical evaluation of a RCT that presents only a CC analysis in which 15% to 30% of outcome measures are missing. The findings would most likely be perceived as potentially biased even when the mechanism of missingness is clearly MCAR as in sample processing errors, a point that has been highlighted in these simulations.

In truth, of course, for a RCT with efficacy levels away from the boundary, participants for whom the (binary) outcome measure is missing contribute nothing more than a collection of baseline characteristics; imputing the missing outcomes does not provide any empirical information about the relation between the exposure and the outcome.

While these findings appear to strengthen the argument in favour of CC analyses, as has been suggested by Liublinska [24], they do not prove that CC methods are necessarily appropriate in all situations. This paper considers just the case of a RCT with a single binary outcome measurement and risk difference estimation (a common practical scenario in malaria studies of efficacy); in more complex study designs (for example longitudinal or cluster randomized studies), and particularly when there are missing observations across many variables, multiple imputation methods remain superior to CC methodologies.

CONCLUSIONS

MI anlyeses must be the primary analyses for the intention-to-treat analyses. These findings provide an argument for the use of the CC approach to always complement MI analyses, with the usual caveat that the validity of the mechanism for missingness be thoroughly discussed. The study also endorses CC methods for per protocol risk difference analyses in these conditions. Pragmatically, while the evidence favour the adoption of a CC analysis when (binary) outcome measures are missing, the reality is that the compromise approach suggested above of a carefully considered primary MI analysis followed by a secondary (essentially sensitivity) CC analysis may be most sensible in terms of getting the findings of such a RCT accepted, even in those situations in which the missing outcomes are "clearly" MCAR or MAR. More importantly, researchers should strive to collect as much data as possible.

Limitations

Different coefficients were used for the treatment group and weight variables to generate different percentage of missing data. Some confounding is thus possible between the impact of including / not including both group membership and weight in the MI analyses and the effect of increasing the percentage of missing data. A better approach might have been to have fixed the effects of treatment group and weight on the missing outcome and then only allowed β_0 to vary to achieve different percentage of missing data. This was an oversight at the study design stage and we are grateful to a reviewer for pointing out this potential interpretation issue. In addition the effect size used for weight was small compared to that used for treatment group; it is possible that this may explain in part why including / excluding treatment effect from the imputation had a considerably greater impact than including / excluding weight.

We also acknowledge that the inflated RMSEs with MI analyses may well be due to the discrepancy between the imputation and analysis models. That fact that logistic regression was used to impute missing outcomes while the OLS regression was used to analyse binary outcome data might impact the size of the RMSEs in the MI analyses.

CONFLICT OF INTEREST DECLARATIONS

Authors declare that there is no conflict of interest/financial disclosure on this work for all authors. We have not been paid any honoraria or otherwise to produce this work.

LIST OF ABBREVIATIONS

CC	Complete Case Analysis
DR-IPW	doubly robust inverse probability weighting
hb	Haemoglobin
H-W	Huber-White
IPW	Inverse probability weighting
ITT	Intention-To-Treat
MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Multiple Imputation
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
OLS	Ordinary Least Squares
OR	Odds Ratio
Para	Parasitaemia
RCT	Randomized Controlled Trial
RD	Risk Difference
RMSE	Root-Mean-Squared Error
RR	Risk Ratio
wt	Weight

AUTHORS' CONTRIBUTIONS

Mavuto Mukaka: Substantial contributions to conception and design, analysis and interpretation of data, drafting the article and the final approval of the version to be published.

Sarah A White: Substantial contributions to conception and design, revising it critically for important intellectual content and the final approval of the version to be published.

Dianne Terlouw: Substantial contributions to conception and design, revising it critically for important intellectual content and the final approval of the version to be published.

Victor Mwapasa: Substantial contributions to conception and design, revising it critically for important intellectual content and the final approval of the version to be published.

Linda Kalilani-Phiri: substantial contributions to conception and design, revising it critically for important intellectual content and the final approval of the version to be published.

Brian Faragher: Substantial contributions to conception and design, interpretation of data, drafting the article, revising it critically for important intellectual content and the final approval of the version to be published.

AUTHORS INFORMATION

Dr Mavuto Mukaka, PhD
Nuffield Department of Medicine, University of Oxford ,UK
Head of Statistics, Clinical Trials Support Group,
Mahidol-Oxford Tropical Medicine Research Unit,
Faculty of Tropical Medicine, Mahidol University, Thailand.

Dr Sarah White, PhD
Senior Lecturer in Biostatistics
College of Medicine, University of Malawi, Malawi

Dr Dianne J Terlouw, MD PhD
Clinical Lecturer/Malaria theme lead MLW;
Liverpool School of Tropical Medicine, Pembroke Place, L3 5QA, Liverpool, UK

Professor Victor Mwapasa, MBBS, MPH, PhD
Professor of Public Health
Dean, Postgraduate Studies and Research
College of Medicine, University of Malawi, Malawi

Dr Linda Kalilani-Phiri, MBBS, MPhil, PhD
Associate Professor of Statistical Epidemiology
College of Medicine, University of Malawi, Malawi

Professor Brian Faragher, PhD
Professor of Medical Statistics, Liverpool School of Tropical Medicine,
Liverpool School of Tropical Medicine, Pembroke Place, L3 5QA, Liverpool, UK

ACKNOWLEDGMENTS:

Thanks to the reviewers comments that helped improve the quality of this manuscript.

Grants and/or financial support: This work was supported by the European and Developing Countries Clinical Trials Partnership (EDCTP) grant number IP.2007.31060.03; and the Johns Hopkins University Center for AIDS Research (Grant Number IP30AI094189) from the National Institute of Allergy And Infectious Diseases.

REFERENCES

1. Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ* 2001; **165**: 1339-1341.
2. Machekano RN, Dorsey G, Hubbard A. Efficacy studies of malaria treatments in Africa: efficient estimation with missing indicators of failure. *Stat Methods Med Res* 2008; **17**: 191-206.
3. Higgins JP, White IR, Wood AM. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clin Trials* 2008; **5**: 225-239.
4. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 2004; **1**: 368-376.
5. Panel on Handling Missing Data in Clinical Trials Committee on National Statistics Division of Behavioral and Social Sciences and Education. The Prevention and Treatment of Missing Data in Clinical Trials Panel on Handling Missing Data in Clinical Trials ; National Research. *The National Academy of Sciences* 2010..
6. Altman DG, Bland JM. Missing data. *BMJ* 2007; **334**: 424.
7. Allison PD. *Missing Data. Quantitative Applications in the Social Sciences*: London, New Delhi., 2001.
8. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; **59**: 1087-1091.
9. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. *Open Med* 2009; **3**: e51-53.
10. Kenward MGJC. Multiple imputation: current perspectives *Stat. Methods Med. Res.* 2007; **16**: 199-218.
11. Ibrahim JG, G. M. Missing data methods in longitudinal studies: a review. *Test. (Madr.)* 2009. ; **18**: 1-43.
12. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med* 2010; **29**: 2920-2931.
13. Magder LS. Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences. . *Control Clin. Trials* 2003; **24**: 411-421.
14. Groenwold RH, Donders AR, Roes KC, Harrell FE, Jr., Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol* 2011; **175**: 210-217.
15. Bell DJ, Nyirongo SK, Mukaka M, Zijlstra EE, Plowe CV, Molyneux ME, Ward SA, Winstanley PA. Sulfadoxine-pyrimethamine-based combinations for malaria: a randomised blinded trial to compare efficacy, safety and selection of resistance in Malawi. *PLoS One* 2008; **3**: e1578.

16. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC. Med. Res. Methodol* 2010; **10**.
17. Royston P, White IR. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software* 2011; **45**: 1-20.
18. Rubin DB. Inference and missing data. . *Biometrika* 1976; **63**: 581-592.
19. Cheung YB. A modified least-squares regression approach to the estimation of risk difference. *Am J Epidemiol* 2007; **166**: 1337-1344.
20. Schafer JL. *Analysis of Incomplete Multivariate Data*: London, 1997.
21. Rubin DB. *Multiple imputation for nonresponse in surveys.* : New York, 1987.
22. White IR, Daniel R, Royston P. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Comput Stat Data Anal* 2010; **54**: 2267-2275.
23. Rezvan PH, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology* 2015; **15**.
24. Liublinska V, Rubin DB. Re: "dealing with missing outcome data in randomized trials and observational studies". *Am J Epidemiol* 2012; **176**: 357-358; author reply 358-359.

Figures

Appendix tables