Comparison of statistical methods of handling missing binary outcome data in randomized controlled trials of efficacy studies

By

Mavuto M.F.J. Mukaka

(BSc, MSc)

Thesis submitted in accordance with the requirements of the University of Liverpool and University of Malawi for the degree of Doctor of Philosophy

July 2013

Certificate of Approval

The Thesis of Mavuto Mukaka is approved by the Thesis Examination Committee:

Prof. Adamson Muula

(Chairman, Post Graduate Committee)

Dr Brian Faragher and Dr Sarah White

(Supervisors)

Prof. Neil French and Associate Prof. Lawrence Kazembe

(Internal Examiners)

Dr Bernard Mbewe

(Head of Department)

Author's declaration

I declare that the work in this thesis was carried out in accordance with the requirements of the University's regulations and Code of Practice for Research Degree programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of others, is indicated as such. Any views expressed in this thesis are those of the author.

NAME: Mavuto M.F.J. Mukaka

SIGNED: 

DATE: 19th July2 013

Acknowledgements

I am indebted to Dr Brian Faragher and Dr Sarah White, my supervisors, who, in spite of their tight programs were able to spare some time to guide and supervise me throughout my studies. They provided wonderful support and advice. I would also like to thank Associate Professor Victor Mwapasa, Associate Professor Linda Kalilani-Phiri and Dr Dianne Terlouw for their mentorship.

I also thank the University of Liverpool and the University of Malawi for enrolling me into a PhD program. I also thank the director and Management of the Malawi-Liverpool-Wellcome Trust Clinical Research Program for allowing me to undertake my studies at this institution. They provided all the support that I needed for my studies.

I am also indebted to the European and Developing Countries Clinical Trials Partnership (EDCTP) for granting me a scholarship to undertake my PhD studies. Your support will always be remembered.

I would further wish to express my sincere appreciation and gratitude to

to my wife Wezi and children Takondwa, Uchizi, Yewo and Tapiwa who missed my social life in most weekends and of course for their encouragement throughout my studies.

Dedication

This thesis is dedicated to my father L.B. Mukaka Juba, my late mother nee T. Mhango (May her soul rest in peace), my wife nee W.Mkandawire and my children: Takondwa, Uchizi, Yewo and Tapiwa.

Contents

List of tables

List of figures

Acronyms/Abbreviations

ACPR	Adequate Clinical and Parasitological Response
AQ	Amodiaquine
ART	Artesunate
CC	Complete Case analysis
CI	Confidence interval
CQ	Choloroquine
DR-IPW	Doubly Robust Inverse Probability Weighting
EC	Extreme case
GEE	Generalized estimating equations
Hb	haemoglobin
HR	Hazard ratio
IPW	Inverse Probability Weighting
ITT	Intention- to- treat
LL	Lower limit of a confidence interval
LOCF	Last Observation Carried Forward
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MICE	Multiple Imputation using Chained quations
MLE	maximum Likelihood estimation
MNAR	Missing Not at Random
NNT	Number needed to treat

OLS	Ordinary Least Squares
OR	Odds ratio
Para	Parasitaemia
PCR	Polymerase chain reaction
RCT	Randomized controlled Trials
RD	Risk difference
RR	Risk ratio
SE	Standard Error
SP	Sulfadoxine Pyrimethamine
UL	Upper Limit of a confidence interval
Wt	weight

Glossary/Definitions

Convergence	Failure by software to provide output or (valid out output)
Covariate	Continuous variable
Efficacy	Proportion of participants with treatment success under controlled conditions expressed as a percentage
Factor	Categorical variable
Missingness	Data being missing
Recrudescence	Retain parasite genotype before and post treatment
Reinfections	New parasite genotype infection after treatment
Substantive model	Data model that is used to estimate parameters of interest

Abstract

The presence of some missing outcomes in randomized studies often complicates the estimation of measures of effect, even in well designed randomized controlled trials. The process may be complicated further when the efficacy rates are close to 0% or 100% as the standard binomial model is susceptible to model non-convergence. The main objective of this study was to compare the performance of multiple imputation (MI) and Complete Case analysis for dealing with missing binary outcomes when modeling a risk difference. Firstly, however, the binomial regression COPY method and the Cheung's modified Ordinary Least Squares (OLS) method were examined using simulation processes for their appropriateness in risk difference modeling. It was found that the number of copies (for the COPY method) required to minimize non-convergence coincided with the number of copies that gave the most biased estimates of the true efficacy difference while increasing the number of copies made the problems of non-convergence and bias worse; using Cheung's method, however, there was 100% convergence with unbiased estimates of effect size. Simulation methods were used to compare the performance of complete case (CC) analysis and several multiple imputation (MI) models for handling missing outcome data over a wide range of efficacy environments and missing value assumptions. When outcomes were missing at random (MAR) or completely at random (MCAR), MI analyses that included treatment group membership in the imputation calculations yielded unbiased estimates of efficacy differences. The CC method was found to be as good, and often better, than MI methods when outcomes were MAR or MCAR, with coverage close to 95% in many situations –

but neither CC nor MI produced unbiased estimates of effect difference when outcomes were missing not at random (MNAR). It was concluded that CC and MI methods are equally good in terms of producing unbiased estimates of effect difference in most missing outcome situations, but applying the intention to treat principle (ITT) which requires all randomized patients to be included in the primary analysis of a RCT, MI should be adopted as the analysis method of first choice, accompanied by a secondary CC analysis for sensitivity purposes (i.e. to investigate the extent of any likely bias).

Chapter 1 : Missing binary outcome data in randomized controlled trials

1.1 Background and motivation

Randomized Controlled Trials (RCTs) are the gold standard for evaluating the impact of treatment or interventions in clinical and epidemiological research (Montori and Guyatt 2001). The most important characteristic of a well designed RCT is that it ensures unbiased estimates of treatment or intervention effect (Montori and Guyatt 2001, Machekano et al. 2008). When the data are fully observed there are well established theoretical methods to derive unbiased estimates of treatment or intervention effect in RCTs. One major requirement made by the use of these standard methods in analysing RCTs is that data are available on all participants recruited in a trial (Allison 2001). In RCTS, most of the baseline data are usually collected and are complete, however, in practice it is very common to find missing outcome data (Wood et al. 2004). Possible reasons for the outcome data to be missing include: loss of participants to follow-up before an outcome of interest is measured; sample processing failure by a piece of laboratory equipment and loss of a participant's laboratory sample (Altman and Bland 2007); participants withdrawing from a study before an outcome of interest is measured due to any of the following: occurrence of adverse events, illness unrelated to a study intervention, protocol violation, and ineffective treatment (Molenberghs and Kenward 2007).

The presence of missing data often complicates analyses and the strength of an RCT design may be compromised. Missing outcome data may lead to increased uncertainty over estimates of treatment effect and biased estimates if not properly dealt with in statistical analyses (Higgins et al. 2008).

For close to four decades now, several statistical methods of handling missing data have been developed with active research still ongoing (Rubin 1976, Dempster et al. 1977, Diggle and Kenward 1994, Robins et al. 1995, Rotnitzky and Robins 1997, Schafer 1997, Scharfstein et al. 1999, Little 2002, Kenward and Carpenter 2007). The methods include the multiple imputation (MI) approach, inverse probability weighting (IPW), doubly robust inverse probability weighting (DR-IPW) and Maximum Likelihood Estimation (MLE). These methods are discussed in detail under literature review in Chapter 2. In spite of the broad body of literature on methods of dealing with missing outcome data, researchers often use the most expedient approach of excluding observations with any missing outcomes which is default in many statistical packages (Allison 2001, Altman and Bland 2007, Machekano et al. 2008). This method is commonly known as complete case (CC) analysis. This CC analysis method may yield biased estimates of treatment effect especially when the missing data levels are high (Donders et al. 2006, Machekano et al. 2008, Altman 2009). Furthermore the CC analysis method lacks a principled statistical foundation and its behavior is unpredictable in different missing data scenarios (Kenward and Carpenter 2007). In some cases researchers choose a method of handling missing outcome data haphazardly during analysis. Choosing a missing data method arbitrarily is dangerous as it ignores

the fact that the existing methods are only valid in specific missing data scenarios. Even the well known principled methods such as the MI methods (methods that fill in the missing observations with randomly generated plausible values based on other observed values) are biased in some situations and are not better than CC analysis in other settings (Allison 2001, White and Carlin 2010). The choice of the methods of handling missing data depends on the pattern of missing data as well as the mechanism that leads to the data being missing (Ibrahim and Molenberghs 2009). The patterns of missing data and missing data mechanisms are detailed in Chapter 2. *In summary*, the methods for handling missing outcome data are well developed for RCTs but are rarely applied in practice (Ibrahim and Molenberghs 2009) and the challenge is on the method choice that is most appropriate for a particular effect measure and missing data scenario since universally robust methods for handling missing data do not exist.

In RCTs, the intervention effect for binary outcomes is often measured using relative risks (RR), odds ratios (OR) or risk differences (RD) (Magder 2003). In recent years an RD has become a widely reported measure of effect in RCTs especially for malaria studies. Examples of trials that report a risk difference include: (Bell et al. 2008, Arinaitwe et al. 2009). Of note, an RD model sometimes fails to converge in software (Cheung 2007).

Simulations are computer intensive procedures that are employed to evaluate the performance of a variety of statistical methods relative to a known value, called a parameter (Burton et al. 2006).

Considering that existing methods are not robust in all missing data scenarios, it is always important to perform simulation studies to compare the performance of different methods of dealing with missing data in order to identify the methods that are the most appropriate for a particular scenario.

Simulation studies have examined methods of handling missing binary outcome data where the summary measure of interest is an OR, for example: (Machekano et al. 2008, White and Carlin 2010, Groenwold et al. 2011). However, little is known on how the missing data methods perform when the outcome of interest is an RD rather than an OR. Clearly there is a gap in our knowledge of the most appropriate methods of handling missing outcome data when estimating the RD from an RCT.

Many principled methods of missing data such as IPW, DR-IPW and MLE are not easy to implement by a general researcher. In contrast, the MI approach is a principled statistical approach for dealing with missing data that is widely available in many software packages and is relatively easier to implement than the other principled methods. Furthermore, MI is valid and efficient in many situations when data are MAR. In spite of its wide availability in statistical software packages, validity, efficiency and relative user-friendliness, researchers rarely apply this method as well as the other principled methods when modeling an RD in the presence of missing outcome data. Perhaps, researchers do not apply this method because its performance has not yet been examined in a simulation study in the context of risk differences. On the other hand, in spite of being adhoc, inefficient and potentially biased, the CC method is a commonly

used method to estimate an RD in the presence of missing outcome data in RCTs. It is important, therefore, to compare the performance of complete case analysis and multiple imputation approach in terms of bias and efficiency, for modeling an RD in the presence of missing outcome data in an RCT setting using simulations over a range of efficacy and missing outcome data scenarios.

Fitting a risk difference model uses the binomial regression model with an identity link function as the standard. However, this analysis approach is susceptible to model fail in software. The Copy method and Cheung's OLS method were potentially identified to be used in cases where the binomial regression model fails.

1.2 The "COPY method" and the binomial regression model

The copy method was proposed by Deddens and Petersen (2003) to deal with the problem of model failure when estimating risk ratios with the log-binomial model using Maximum Likelihood Estimation (MLE). The non-convergence often arises when either or both of the individual risk estimates is close to either 0% or 100%, so the ratio itself is either close to zero or approaches infinity). In this approach, multiple copies of the dataset are added to the original set; when the binomial regression model is applied to this modified data set, the model converges and approximate maximum likelihood estimates of the risk ratio are obtained (Deddens and Petersen 2003, Deddens and Petersen 2008, Petersen and Deddens 2009).

Mathematically, the copy method calculates MLEs using a log-binomial model on an expanded version of the data set that contains K-1 copies of the original dataset plus one copy of the original dataset in which the values of the binary outcome variable are reversed (the 1's (successes) are all changed to 0's (fails) and the 0's (fails) are all changed to 1's (successes)). When modeling a risk ratio using a log-binomial regression model, if the total number of dataset copies, K, is finite, the iterative estimation solution moves away from the parameter space and is an MLE for the "copied" dataset (Petersen and Deddens 2009).

As K gets increases, the MLE estimate obtained from the "copied" dataset with a log-binomial model approaches the MLE estimate for the original dataset (i.e. is asymptotic) (Deddens and Petersen 2008, Petersen and Deddens 2008, Petersen and Deddens 2009). Petersen and Deddens (2008, 2009) recommend that K should be at least 100 (although in their paper they used a value of K = 1,000).

Mathematically, expanding the original data set in the manner required for the copy method is simply equivalent to creating a new data set consisting of one copy of the original data set having a weight of K-1 and one copy of the original data set with the outcome values reversed having a weight of one. Lumley (2006) states that use of the weights (K-1)/K and 1/K for the original outcome and the reversed outcome datasets respectively eliminates the need to adjust the standard error (Lumley et al. 2006). The COPY method was examined to assess whether it is an appropriate alternative when the standard binomial model fails. It was investigated in terms o f convergence and bias.

1.3 Cheung's modified OLS method

Cheung (2007) proposed a variation of the Ordinary Least Squares estimation methodology to address the address the problem of non-convergence when modeling risk difference. The method uses a modified least-squares regression approach with a Huber-White robust standard error (Cheung 2007).

Simulation studies were performed to investigate the suitability of this method for modeling risk differences, in terms of both convergence and bias as an alternative to the binomial regression.

1.4 Description of the motivating malaria efficacy clinical trial data

1.4.1 Study design

This research was motivated by a malaria efficacy study that was conducted in Malawi between 2003 and 2005 in children aged 1 to 5 years. The methods and the findings of this study are detailed in Bell et al (2008) but a brief summary of its rationale, design and findings follow:

Bell and colleagues conducted a blinded randomized controlled study to compare the efficacy of several Sulfadoxine-Pyrimethamine (SP)–Based Combination therapies. They used a total of four treatment groups including the placebo group. The three "active" (comparator) treatments were SP plus chloroquine (CQ), SP plus artesunate (ART) and SP plus amodiaquine (AQ); the control arm comprised of SP plus placebo

(SP was the standard first line treatment for uncomplicated falciparum malaria in Malawi during the time when the study was conducted). The study was conducted in response to accumulating evidence indicating that SP was developing some resistance and because the WHO was recommending the use of combination therapies (especially the artemisinin combination therapies (ACTs)).

The study was done in Malawi and was based at the Chileka Health Centre. Chileka is a rural area in southern Malawi that has perennial malaria transmission that peaks during the rainy season. The rainy season in this area is between October and May, with the heaviest rains occurring somewhere around December to April.

All children in the study area aged between 1 and 5 years were screened for uncomplicated falciparum malaria. Children were recruited into the RCT if their weight was greater than or equal to 6 kg, if they had an axillary temperature of greater than or equal to 37.5^0C, if they had not been treated with an antimalarial drug or cotrimoxazole in the previous 4 weeks, and if they had a plasmodium falciparum parasite density of between 2000 and 200,000 parasites/ml – but were excluded from recruitment if they had any signs of severe malaria.

Children who met the inclusion criteria were enrolled and randomized in blocks of 12 (i.e. with 3 children allocated to each of the four treatment arms in each block). Each child was assigned a randomization (study) number sequentially.

Measurements were taken from each recruited child on days 0, 1, 2, 3, 7, 14, 28 and 42 and on any other unscheduled day if they were sick during follow up. The clinical outcomes were assessed according to the 2003 WHO efficacy protocol (World Health Organization 2003). Children were withdrawn from the study if they missed a follow up visit detailed above, if they (or their guardian/carer) withdrew consent or if they took treatment that was considered to be a protocol violation/deviation.

1.4.2 Sample size and statistical methods

The study was designed to have 90% power to detect the following difference in the proportion of children with an ‘‘adequate clinical and parasitological response’’ (ACPR) efficacy rate at the conventional 5% significance (alpha) level: 80% response in the SP plus placebo arm vs. 95% in at least one of the combination therapies. The literature available at the time the study was being designed indicated that SP was developing high resistance, so the efficacy of this treatment arm was anticipated to be 80% or lower. It was planned that each of the combination therapies would be compared in turn with the SP plus placebo group.

It was estimated that 85 children would be required in each treatment arm (total 340 children) to detect the desired combination treatment effect size. To allow for a loss to follow-up rate of up to 15%, the actual sample size was set at 100 children per treatment arm (total 400 children).

The primary endpoint was the day 28 ACPR rate (i.e. the proportion of children who had an ACPR by day 28). Day 28 ACPR rate was thus a binary variable indicating

whether each participating child had a treatment failure or treatment success. Children were said to have had a treatment success if they had no fever and no parasaitaemia, otherwise they were said to have had a treatment failure.

The primary analysis of the primary endpoint was conducted using the intention to treat (ITT) principle whereby all children recruited into the trial were included in the analyses according to the group that they were randomized to. This was followed by a secondary analysis conducted using a per protocol analysis strategy (as a form of sensitivity analysis).

Almost all values were available for the baseline variables because these formed part of the inclusion or exclusion criteria. However, there were some children with missing outcomes due to a number of reasons, including: lost to follow-up, withdrawal of consent, withdrawn from study because of a protocol violation or deviation before the day 28 outcome was assessed. A few children had mixed plasmodium falciparum malaria genotypes post treatment. In many children for whom parasitaemia was detected during follow up, PCR was used to determine whether treatment was a success or not; however the outcome remained indeterminate for some children even after performing PCR.

The missing outcomes created some challenges in terms of how they should be treated in the statistical analysis. In the intention to treat analyses, all children with missing outcomes were all classified as successes in a first analysis and then as failures in a

second analysis irrespective of which treatment arm they had been allocated to. In the per-protocol (PP) analysis, all children with missing outcomes were excluded from the analytical process (again irrespective of which treatment arm they had been allocated to) - a method that is commonly referred to as "complete case analysis". In order to improve accuracy, the PP analyses were done using polymerase chain reaction (PCR) corrected data to distinguish recrudescences from reinfections. When PCR results showed that a post-treatment parasitaemia genotype was a reinfection, the outcome was classified as a treatment success to the original parasitaemia genotype. If PCR result was indeterminate, the child was assigned a missing outcome value and was excluded from the per protocol analysis.

All data analyses were performed using the Stata for Windows software (version SE/8; statacorp; College Station, Texas 77845 USA). The outcome statistic chosen to indicate effect size was the risk (efficacy) difference between each intervention group and the control arm. Binomial regression models were fitted to the data and used to estimate the relevant risk differences (along with their corresponding 95%confidence intervals).

1.4.3 Missing outcomes

By day 14 of the study, 44 (9.7%) of all children recruited had been withdrawn, and this number had risen to 51 (11.2%) by day 28. Consequently these children had a missing primary endpoint on day 28. The reasons for withdrawal included: lost to follow-up; protocol violation; vomiting medication on the first day: voluntary withdrawal of

consent. The proportions of children with missing outcomes were similar across the treatment groups. Only negligible missing levels were observed in the covariates.

1.4.4 Results for the primary outcome of the historical data: efficacy of antimalarials

The efficacy rates were very high in all of the intervention arms. The AQ plus SP had an efficacy rate that was as high as 97%, 95% CI (93%, 99%) by day 28. The efficacy rate was close to the boundary value of 100%. Using the ITT analysis strategy in which the missing outcomes were assigned success values, the day 28 ACPR rate was lowest in the SP plus placebo group, which had an efficacy rate of 25%, 95% CI (18%, 34%), much lower than that anticipated (80%). The AQ+SP group had an ACPR rate of 97%; this was significantly higher than for the CQ+SP and ART+SP groups which had efficacies of 81%, 95% CI (73%, 88%) and 70%, 95% CI (61%, 78%) respectively (thus proving that in malaria treatment studies, efficacies close to the boundaries are just as possible as efficacy levels that are away from the boundary). There was no significant difference between the CQ+SP and ART+SP groups.

1.4.5 Motivation for designing simulation studies of missing data methods

During the analysis of this data, we (the co-investigators) noted that there was no clear guidance on the choice of an appropriate method of handing missing binary outcome data arising from a randomized controlled design when risk differences are of interest.

From the literature on similar studies, we noted that authors tended to adopt a method of handling missing data arbitrarily without justification. Most commonly, methods were selected adhoc and usually involved extreme case (EC) analyses and complete case (CC) analyses. The CC analysis simply excludes any cases with missing outcome data while EC analysis simply replaces missing outcomes either all as successes or all as failures. Both of these methods are prone to bias especially when the levels of missing data are high. Statistical power may also be reduced in the case of complete case analysis.

In the absence of a clear guidance we were tempted to just choose the missing data methods that were being commonly used and to adopt the adhoc methods of complete case analysis for the per protocol analyses and extreme case analysis for the intention to treat strategy. We did not have any clear basis for the choice of these methods apart from being consistent with other researchers who had reported on the same subject area.

This rather negative experience was the motivation to start to think of carrying out a simulation study that would provide empirically based guidance on the choice of methods for handing missing outcome data. We noticed that there were a number of principled approaches published for handling missing data in such scenarios but the main challenge was on the method choice.

We were aware that the efficacy rates in the Bell et al study were very variable- some were very close to the 100% boundary while others were some distance away from the boundary – which could also complicate the analyses. For example, efficacies rates of 25%, 70%, 81% and 97% were observed in the "SP plus placebo", "ART plus SP", "CQ plus SP" and "AQ plus SP" treatment arms respectively. This provided a rationale for the choice of efficacy levels to be considered in a simulation study; efficacy levels were chosen in such a way that they covered the whole of the expected efficacy spectrum – some of the efficacy levels were chosen to be close to boundary values (in this context, close to 100%) while others were chosen to be away from the boundaries.

In addition to the Bell et al study, another malaria efficacy study was underway in Malawi at the time of formulating this dissertation project. So, in order to both inform a more informative analysis of the Bell et al study and to guide the preparations for the analyses of this second study, simulation studies were planned to address a clearly identified gap in our knowledge about the optimum choice of the methods of handling missing outcome data (with particular emphasis given to the situation of a binary outcome measure). The missing data simulations were planned to be performed over a wide range of efficacy rates and over a range of assumption of the mechanisms that may be reacting the missing outcomes. The simulation studies on missing data methods are the core of this thesis.

In most comparative studies, adjusted estimates of treatment effect are of interest, both to identify factors that are independently associated with the treatment effect size and to ensure that the effect size estimate presented is a true indication of the effect of the treatment of interest alone. However, risk difference modeling using the standard binomial regression model is susceptible to model failure (model convergence problems and estimate bias) when adjusting for other variables. This phenomenon provided additional motivation to examine factors that may be associated with the model failure and also to perform simulation studies to identify alternative methods that can be used when the standard binomial regression model fails to provide an adjusted estimate of effect size.

In summary, the analysis of this historical data was very motivating. It was clearly observed that although methods of handling missing data are well developed, there is a gap in knowledge of the methods that are the most appropriate for modeling a risk difference in the presence of missing binary outcome data in the context of a randomized controlled design. Furthermore, no universally robust methods of missing data exist. Thus, the primary rational for the simulations presented in this thesis was to compare methods of handling missing data and thereby to identify the most robust method of analysis. In addition, it was deemed important to perform simulation studies to understand the issue of non-convergence when modeling a risk difference using the standard binomial model and to identify alternative methods for dealing with this problem.

1.5 Aims of the project

1.5.1 Main objectives of the study

The primary objective of this research is to use computer simulation techniques to compare the performance, in terms of bias and efficiency, of the MI method and CC analysis approach for handling missing binary outcome data when modeling a risk difference in the presence of missing outcome data. The comparisons are performed over a variety of efficacy and missing binary outcome data scenarios, focusing on a randomized controlled trial design. Furthermore, this study identifies the factors that may lead to the convergence problems in software when estimating an adjusted risk difference.

1.5.2 Specific objectives

1. To compare the performance of the multiple imputation (MI) method and complete case (CC) analysis when estimating a risk difference from a randomized controlled design with the following missing data mechanisms:
    a. missing at random(MAR);
    b. missing completely at random(MCAR);
    c. missing not at random (MNAR).
2. To assess how the closeness of the efficacy to a boundary value (i.e. to 0% or 100%) impacts the estimates from CC and MI in terms of bias and efficacy by considering the following efficacy scenarios:
    a. the efficacy rates are away from a boundary value in both arms;

b. the efficacy rate in one treatment arm is close to a boundary value while in the other arm it is away from a boundary value;

c. the efficacy rates are close to a boundary value in both treatment arms.

3. To investigate the impact of the following factors on non-convergence of a binomial model that estimates a risk difference in Stata statistical software package:

   a. one or both efficacy rates close to a boundary value;

   b. the number of covariates in a model;

   c. the correlations between covariates.

4. To assess the appropriateness of the binomial model, the COPY method of the binomial model and Cheung's OLS method in terms of convergence and bias.

1.6 Significance of the study

"Risk difference" is becoming a commonly used measure of effect in medical research especially in randomized controlled trials. However, while a risk difference model may lead to biased estimates of intervention effect in the presence of missing outcome data, the exact nature of this problem is not fully understood. The findings of this research will fill this gap in knowledge by providing informed guidance on the relative merits of the MI method and CC analysis approach for handling missing binary outcome data under different missing data mechanisms and efficacy scenarios. Knowledge of the factors that may lead to a risk difference model failure will help researchers in deciding whether to use CC or MI methods and, if the latter are chosen, which covariates should

be used in the MI process that are most likely to produce unbiased estimates of the effect difference and minimize the risk of model failure.

1.7 Structure of the thesis

The thesis is structured as follows: Chapter 2 discusses the literature on the common measures of effect for binary outcome in randomized controlled trials, missing data theory and common approaches for handling missing data. Chapter 3 provides details of the statistical and simulation methods compared in this dissertation. Experimental findings of the simulation studies on convergence of a binomial model and alternative approaches are presented in Chapter 4. Chapter 5 presents experimental findings of the simulation studies for the comparisons of methods for dealing with missing data. The discussion of all experimental findings is presented in Chapter 6. Study conclusions, recommendations and further research questions are also presented in Chapter 6. The Stata programs used in the simulation exercises are provided as an appendix at the end of the thesis.

Chapter 2 : Literature review

This Chapter describes the common measures of effect for a binary outcome variable in RCTs and reviews missing data theory in this situation, focusing on common approaches to handling missing binary outcome observations. Both unprincipled and principled statistical methods are reviewed, and the strengths and weaknesses of each are discussed.

This Chapter is structured as follows: firstly, common measures of effect when modeling a binary outcome data in RTCs are described; missing data theory in general is then reviewed; finally, common approaches to handling missing binary outcome observations are discussed.

2.1. Common measures of effect for binary outcome data in RCTs

The measures of effect in RCTs where the outcome of interest is binary include: odds ratios (OR), relative risk/risk (rate, hazard) ratios (RR), risk (rate, hazard) differences (RD) and number needed to treat (NNT). However, the most commonly used measures are: OR, RR and RDs; there are many examples of these measures reported in the research literature for example: (Faucett et al. 2002, Brasseur et al. 2007, Bell et al. 2008, Borrmann et al. 2008, Crompton et al. 2008, Faucher et al. 2009, Gesase et al. 2009, Chasela et al. 2010, French et al. 2010).

For mainly historical reasons, ORs are the most widely reported measure of effect in clinical and epidemiological research, including RCTs. Until fairly recently, there were

computational difficulties such that regression models of binary data could only be done on odds ratios and not on risk or rate ratios, even though the theory had been well developed (Cox and Snell 1970). In fact the Cox proportional hazards regression model (Cox 1972) was the first method adopted widely to model rate ratios before Poisson and negative regression models appeared in the widely used statistical computer packages. The Cox proportional hazards regression model estimates risk ratios by setting the follow-up time to 1 (Cummings 2009b).

However, there has been a lot of debate as to which of a RR or an OR is the most appropriate statistic to use (Barros and Hirakata 2003); this debate has been extended to also include RDs. It has been argued that the OR has been in wide use because it is computationally less challenging than the alternatives RR and RDs (Wacholder 1986). When analyzing case-control studies, of course, the OR is the only appropriate statistic to use to compare risks; the relative risk is mathematically invalid because the selection of study participants is based on outcome and not exposure (Miettinen and Cook 1981).

Despite their computational advantages and frequency of use in the past, the continued use of ORs for other than case-control studies has been heavily criticized. An OR is practically challenging to interpret for many medical researchers (Greenland 1987, Barros and Hirakata 2003). It is often interpreted as an approximation of a relative risk when in fact this approximation is only valid when the outcome of interest is not common (Greenland 1987, Barros and Hirakata 2003). The main concern among epidemiological researchers is that when an OR is misinterpreted as a relative risk even when the

prevalence of the outcome measure is common, can easily yield incorrect inferences of treatment effect (Davies et al. 1998, Case et al. 2002, McNutt et al. 2003, Page and Attia 2003, Cheung 2007).

An OR is routinely chosen as a measure of effect often based on mathematical convenience without consideration of whether the results are interpretable (Walter 2000). The relative risk, on the other hand, occupies a very special role as a measure of effect for binary outcome in cohort studies, in which exposure usually precedes outcome and it is generally befitting to use an RR as a measure of effect. For some time now, many researchers have been recommending a systematic use of the RR rather than the OR whenever appropriate (Axelson et al. 1994, Davies et al. 1998, Grimes and Schulz 2008, Cummings 2009a). RDs and RRs are sometimes more biologically plausible than ORs which may give them an advantage when describing a risk (Walter 2000).

Further to this debate, Cheung (2007) points out that in the case of equivalence and non-inferiority studies, the RD has a more meaningful interpretation than either the OR or the RR (Cheung 2007). Unfortunately, however, much as the risk difference is becoming a more attractive measure of effect for binary outcome measures than both the OR and RR because of its interpretational advantages, as will be discussed later in this dissertation, the binomial mathematical models used to compute the RD often suffer from convergence problems in many (possibly all) statistical software packages (Wacholder 1986, Cheung 2007).

2.2 Risk difference modeling and alternative methods

2.2.1 Odds ratios and risk ratios

In comparative studies where the outcome of interest is binary, odds ratios or risk ratios are most commonly used to model intervention effect size.

- The odds ratio is the only valid summary statistic in case control studies where the selection of study participants (cases and controls) is based on the outcome rather than the exposure - in this case a risk ratio cannot be computed directly. More precisely: the odds ratio compares the relative odds of a (binary) outcome between treatment groups; it is also a commonly used measure of strength of association between outcome and exposure in case-control studies. Odds ratios can most easily be obtained from a logistic regression model.
- The risk ratio is the most appropriate summary statistic in cohort studies where exposure usually precedes outcome and study participants are selected on the basis of their exposure to the risk factor of interest (then followed up to determine the incidence of the outcome of interest). More precisely: the risk ratio compares directly the probability of a dichotomous/binary outcome between two groups and is a standard measure of effect in cohort studies when the objective is to compare (binary) outcomes between groups. Rate ratios can most easily be obtained from a negative binomial regression model.

The odds ratio is still used widely as the preferred measure of effect even in situations where a risk ratio is probably more appropriate. There are several reasons for this:

- The odds ratio has important computational advantages over both the risk difference and the risk ratio when adjusting for covariates (Wacholder 1986).
- The odds ratio does not have the convergence problems encountered in many software packages when attempting to estimate and/or manipulate risk ratios and risk differences.
- Most reports of cohort studies in the past have used odds ratios, so using this statistic for new cohort studies provides a more ready comparison with findings from previous studies.

On the debit side, however, the odds ratio is often used – and interpreted – as if it is actually a risk ratio (Zhang and Yu 1998). Such an interpretation is only valid when the outcome of interest is rare (Greenland 1987, Zhang and Yu 1998). Odds ratios are not the simple concept many believe them to be (Case et al. 2002, Cheung 2007), so are sometimes misinterpreted. Predominantly, many researchers interpret the odds ratio as if it is actually a risk ratio (Davies et al. 1998, Case et al. 2002). Such an interpretation may lead to an incorrect inference of the treatment effect estimate (Case et al. 2002). The degree of error in interpreting an odds ratio as a risk ratio is often small (Davies et al. 1998, Page and Attia 2003), but can be substantial in some situations (McNutt et al. 2003, Page and Attia 2003). The extent of any interpretive error cannot be readily assessed

from the value of the odds ratio estimate as this error is a function of the underlying probabilities rather than of the odds ratio itself.

The risk ratio also has some limitations, including potentially important interpretation problems. For example, the risk ratio for the outcome Y = 0 is not the inverse of the risk ratio for the outcome Y= 1 (Blizzard and Hosmer 2006). This means that when risk ratios are used in some RTCs and equivalence studies, evidence of equivalence in the failure rate does not necessarily come with evidence of equivalence in the success rate (Cheung 2007). This dilemma considerably limits the use of the risk ratio in some situations (Cheung 2007).

2.2.2 Risk difference and rationale for its choice

The risk difference is an alternative statistic for presenting effect size estimates derived from binary outcome data, and seems to be the most appropriate method in situations where efficacy is high in both treatment groups.

For example, consider a trial that recruits 1000 individuals in the intervention arm and 1000 participants in the control arm. Let the treatment failure rate be 4% (40/1000) in the control group and 1% (10/1000) in the intervention group (i.e. the success rates are 96% and 99% in the control and intervention arms respectively). This gives a RR of 4.0 (95% CI 2.0 : 8.0; $p<0.001$). The interpretation of this is that individuals in the control group

were 4 times (and statistically significantly) more likely to fail than those in the intervention group, with the 95% confidence interval indicating that the true risk ratio lies between 2 and 8. On the other hand, using the odds ratios as an alternative for this data gives OR=4.1 (95% CI 2.0 : 9.3; $p<0.001$). The interpretation of this is that the odds of treatment failure in the control group is just fractionally over 4 times (and statistically significantly) greater than the odds of treatment failure in the intervention group and that we can be 95% confident that the true population odds ratio lies somewhere between 2.0 and 9.0. Clearly, both of these statistics considerably exaggerate the real (i.e. clinically important) difference in the relative effects of the two treatments.

Now consider the analysis of this study using risk differences. This would give RD=0.03(95% CI 0.02 : 0.04; $p<0.001$). The interpretation of this statistic is that there is a (statistically significant) 3% difference in the risk of treatment failure between the intervention and control arms, and that we can be 95% confident that the true population risk difference is between 2% and 4%. In this situation, the risk difference is clearly providing a more sensible and meaningful assessment of the relative sizes of the treatment failure rates than either the odds ratio or risk ratio. It is common in malaria efficacy studies to observe treatment efficacy rates that are greater than 90% and just slight differences in the success (or failure) rates in such studies are likely to be exaggerated if presented as an odds ratio or a risk ratio.

As studies are increasingly being analysed using risk differences, it is becoming appealing for new studies to be analysed in a similar manner so that the outcome measures from different studies can be combined in a meta-analysis /systematic review. Risk difference estimation is preferred by many researchers because it is easier to interpret than the alternative odds ratio. A risk difference is symmetric so evidence of equivalence in failure rate is mathematically equivalent to evidence of equivalence in success rate (Cheung 2007); risk ratios, on the other hand, are not symmetric. These trends were the primary motivating factor for consider a risk difference model in the simulations presented in this dissertation.

The binomial regression model with an identity link function is used to fit a risk difference model. However, fitting a risk difference model often encounters the problem that the binomial regression model fails to converge (Cheung 2007).

2.2.3 Mathematical principles underlying analytical approaches in risk difference models

This section describes the mathematical principles underlying the logistic regression and risk difference models. Firstly the logistic regression model is considered, thereafter a risk difference model is described.

Estimation of probabilities a logistic regression model (Odds ratios)

Consider the following generalized linear model (GLM) again:

Replace μ with π to estimate probabilities:

$$g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{.....(2.1)}$$

$$g(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{.....(2.2)}$$

If the aim of the analysis is to estimate odds ratios, a logit link function is used in the GLM and the GLM becomes:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{.....(2.3)}$$

$$\pi = \frac{\exp(\alpha + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + ... + \beta_k X_k)} \quad \text{.....(2.4)}$$

$$\pi = \frac{1}{1 + \exp(-(\alpha + \beta_1 X_1 + ... + \beta_k X_k))} \quad \text{.....(2.5)}$$

This will result in probability estimates that lie between 0 and 1. This is the reason why logistic regression model is likely to converge and provide sensible estimates from the equation 2.6 below:

$$\alpha + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{.....(2.6)}$$

Estimation of probabilities from risk difference model (binomial regression with identity link)

Consider the following generalized linear model (GLM)

$$g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

where g(u) is a link function that identifies a function of the mean that is a linear function of the covariates and $X_1 \cdots X_k$ is a set of k explanatory variables.

When the outcome is binary, u becomes π (the proportion of participants with an outcome of interest); in other words, π is the probability of observing a specific category of the binary outcome. Thus, the GLM can be re-written as in equation 2.2:

$$g(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

If the aim of the analysis is to estimate risk differences, an identity link function is used in the GLM which reduces to:

$$\pi = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{.....(2.7)}$$

From the equation above, it should be noted that the estimate of π as a linear function of explanatory variables can easily yield estimates of probabilities that are outside the valid range 0 to 1. This is so because the expression $\alpha + \beta_1 X_1 + \cdots + \beta_k X_k$ is unbounded and can yield values that range from -∞ to +∞. But since a binomial model is constrained to estimate probabilities that are between 0 and 1, the estimates of probabilities that are outside this range may result in computer software not providing results.

Now if we consider (2.7) $\pi = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$

and suppose X1 is a binary exposure (0 or 1) that may denote the treatment/ intervention that an individual is assigned to. Then the estimate of the adjusted risk difference becomes:

$$RD = \hat{\pi}_1 - \hat{\pi}_0 \quad \text{.....(2.8)}$$

$$RD = \{(\hat{\alpha} + \hat{\beta}_1 \times 1 + \cdots + \hat{\beta}_k X_k) - (\hat{\alpha} + \hat{\beta}_1 \times 0 + \ldots + \hat{\beta}_k X_k)\} \quad \text{.....(2.9)}$$

$$RD = \hat{\beta}_1 \quad \text{.....(2.10)}$$

So the estimate of the risk difference is just $\hat{\beta}_1$ and this does not have boundary constraints as is the case with the estimation of probabilities. This suggests that if interest is in estimating the risk difference rather than the individual risks (probabilities) themselves, estimates of the risk difference based on the above linear model would be valid. This is also demonstrated by Cheung (2008), when he established that Ordinary Least Squares estimation methods with Huber White standard errors are valid for the estimation of risk differences. This method also avoids the non-convergence problems that can be experienced when using the binomial regression model with an identity link function because the core function of this binomial regression model is the estimation of probabilities. It has already been shown above that such estimation of probabilities based on the standard binomial regression model may result in probabilities that are outside 0 and 1 because the linear function of the covariates is unbounded. The Cheung's method for modelling risk difference is described in detail in section 3.13 of this dissertation.

2.3 Theory of Missing Data

In this section missing data theory is reviewed.

2.3.3 Missing Data Mechanisms

Missing data is a common problem in many research disciplines. The process that results in missing data is technically known as the *missing data mechanism* (Rubin 1976, Schafer 1997, Little 2002). When handling missing data in a statistical analysis, the choice of missing data methods hugely depends on the missing data mechanism (Schafer 1997). In order to choose an appropriate statistical methodology, therefore, it is imperative for the analyst to have an idea or some plausible assumption of the missing data mechanism present in the data. In his theory, Rubin (1976) developed a useful taxonomy for describing the assumptions regarding the missing data mechanisms which provides an important guide to researchers on how to deal with missing data in analyses. The missing data theory regards the missingness of data as a probabilistic event. The commonly made assumptions about the distribution of missingness and the missing data taxonomy are based on Rubin (1976) are reviewed below.

2.3.4 The Distribution of Missingness

In this section, the general notation and distribution of missingness is described for a general longitudinal study where outcome is repeatedly collected over time. Thereafter notation and distribution of missingness is described for a special case where outcome is measured only once- at the end of the study.

Consider a longitudinal study into which n subjects are enrolled and followed up over time such that measurements are taken at baseline and then at specified times during the follow up (t assessment times in total).

Let Y_{ij} and X_{ij} represent the outcome and a covariate data respectively for study participant i (i = 1,2,… , n) measured at time j (j = 1,2,…, t).

In general Y_{ij} and X_{ij} denote a complete dataset for a longitudinal study where data is collected repeatedly over time. In a typical RCT study, for reasons that may be beyond the investigators' control, not all of the Y_{ij} and X_{ij} will be observable for all study participants no matter how rigorous the researchers may be (i.e. there will be missing data, and these may occur in the outcome variable as well as in the covariates). The reasons for missing data are usually difficult to precisely determine in practice.

Let $Y_i = (Y_{i1},...Y_{it})^T$ be a complete vector of outcomes for individual *i* taken at times $j = 1,2,....,t$

This data can be partitioned into those time points at which the outcome has been observed and those whose time points at which the outcome is missing.

Thus Y_i can be re-written as $Y_i = \left(Y_i^{(obs)}, Y_i^{(miss)}\right)$, where $Y_i^{(obs)}$ denotes the observed data and $Y_i^{(miss)}$ denotes the missing outcomes which should have been observed for individual i.

Further, let $D_i = \left(D_{i1},...D_{it}\right)^T$ be a vector that indicates missingness such that $D_{ij} = 1$ if $Y_{ij} \in Y_i^{(miss)}$ and $D_{ij} = 0$ if $Y_{ij} \in Y_i^{(obs)}$ where ϵ means 'belongs to'

Since, as has been previously mentioned, it is often difficult to precisely establish the source of missing data in a dataset, the probability distribution function of the missing data indicator variables D_i given the fully observed data Y_i, is often used as the best statistical tool to explain the process that is creating missing data in a dataset and is denoted as $f(D_i \mid Y_i)$ (Schafer and Graham 2002).

In this thesis, a special case where missingness is only in the outcome and where the outcome is measured only once at the end of the study is considered. Furthermore, the baseline variables are collected at the beginning of the study only. This is a common design in malaria efficacy RCTs. Now considering this special case, the notations and descriptions are presented as follows:

Let Y_i and X_i represent the outcome and a covariate respectively for study participant i (i = 1,2,… , n) taken only at one time point.

Let Y be a complete vector of outcomes for all individuals in the study. This data can be partitioned into those whose outcomes have been observed and those whose outcomes are missing.

Thus, Y can be presented as $Y = \left(Y^{(obs)}, Y^{(miss)}\right)$, where $Y^{(obs)}$ denotes the observed binary outcome data and $Y^{(miss)}$ denotes the missing binary outcomes which should have been observed.

In this special case, let D be a vector that indicates missingness such that $D = 1$ if $Y_i \in Y^{(miss)}$ and $D = 0$ if $Y_i \in Y^{(obs)}$ where ϵ means 'belongs to'.

The standard nomenclature for missing data mechanisms (Rubin D.B 1976) classifies missing data mechanisms as (i) missing at random (MAR), (ii) missing completely at random (MCAR), or (iii) missing not at random (MNAR).

2.3.5 Taxonomy for missing data mechanisms

This section firstly presents taxonomy for missing data for a general longitudinal study where outcomes are repeatedly collected over time. Thereafter taxonomy is described in terms of a special case where missing data is only in one outcome variable and the outcome is measured only once at the end of the study as this is the scenario that is considered in the study simulations that have been considered in this thesis

2.3.6 Missing At Random (MAR)

In longitudinal studies where data is repeatedly collected overtime, the outcome data are said to be *missing at random* (MAR) if the probability of being missing depends on the observed outcome values and the covariates, but is independent of the specific missing outcome values that should have been observed in principle. In mathematical terms this is expressed as follows:

$$f(D_i \mid Y_i, X_i, \eta) = f(D_i \mid Y_i^{(obs)}, X_i, \eta) \quad \text{for all } Y_i^{(miss)}, \ \eta \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.11)$$

where η denotes a set of unknown parameters governing the missing data indicators; Y_i , X_i and D_i are the complete outcome data, the covariates and missing data vectors respectively.

Example

Consider a study in which age (years) and haemoglobin level (Hb) in g/dl are two variables that a researcher intends to observe on each participant. For now, consider only those study participants whose age is observed and equal to a particular value, say 10 years old. In practice Hb may be missing for some of these participants aged 10 years. MAR implies that among these study participants with observed age = 10 years, the distribution of Hb values is the same among the cases for which Hb values have been observed as it is among the cases for which Hb level is missing. Similarly, for study participants with observed age = 11 years, the distribution of Hb is the same among the

cases for which Hb is observed as it is among the cases for which Hb level is missing; however, the distribution of Hb values for participants with observed age = 11 may be different from the distribution of HB values for those with observed age = 10 years. The same applies for participants with observed age = 12 years, 13 years… etc. That is, the missing Hb values should be regarded as a random sample of all the Hb values within the observed age subgroups.

The practical challenge with the MAR mechanism is that it is often difficult to confirm that the probability of data on Y being missing entirely as a function of observed data (Little 2002). However, MAR is often plausible in practice (Schafer and Graham 2002, Kenward and Carpenter 2007).

2.3.7 Missing Completely At Random (MCAR)

The outcome data are said to be *missing completely at random* (MCAR) if the probability of being missing does not depend on either the value of the outcome Y or the value of the covariate.

Mathematically this is expressed as follows:

$$f(D_i \mid Y_i, X_i, \eta) = f(D_i \mid \eta) \text{ for all } Y_i, X_i, \eta \quad \text{……………………………………… (2.12)}$$

In fact this is a special case of MAR in which the probability that data are missing is independent of both the specific missing values that in principle should have been observed and the values of the observed data (Schafer 1997, Schafer and Graham 2002).

Example

Revisiting the example above, MCAR implies that the missing Hb values are neither related to age nor to the other Hb values (whether observed or not). Thus, with MCAR, the component of the data with missing Hb values is a random subset of the complete original sample of Hb values – and equally, by definition, the observed sample is also a random sample of the original complete sample. In fact, MCAR implies that the missing values are a random sample of all values of the population from which the study sample came (Rubin 1976). A typical practical example of MCAR would be a test tube containing a laboratory sample being accidentally broken before the sample has been processed or the sample become contaminated or non-viable because of an electricity failure to the storage facility.

2.3.8 Missing Not At Random (MNAR)

The outcome data are said to be *missing not at random* (MNAR) if the probability that the outcome data are missing depends on both the observed outcome values and the unobserved outcome values. In mathematical terms this is expressed as follows:

$$f(D_i \mid Y_i, \mathrm{X}_i, \eta) = f(D \mid Y^{(obs)}, Y^{(miss)}, X_i, \eta) \text{ for all } Y_i, X_i, \eta \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.13)$$

Example

Consider an HIV longitudinal study in which CD4 count is measured for each participant at each clinic visit during follow-ups. Suppose that some participants ~~may~~ drop out of the study before the study ends due to HIV related death. These "drop-outs" will have missing CD4 count at visits scheduled after they died; it is also likely that they will also have low CD4 counts as this is known to be related to HIV related death. In such cases, therefore, the missing CD4 counts will be correlated with the actual missing values (i.e. those with a low CD4 are more likely to die due to HIV related death than those with a high CD4, or more relevantly to the context of this dissertation, those with unobserved but a low CD4 count at a missing visit will be more likely to have a missing CD4 count value at the that visit because of the value of the CD4 itself which is low and therefore may be related to death from HIV related cause). This would be a typical example of MNAR.

If data are NMAR then the missingness mechanism is referred to as non-ignorable. When the missing data are non-ignorable the missingness models should be correctly specified in order to obtain consistent estimates of the parameters of interest (Diggle and Kenward 1994, Little 2002).

2.3.9 Taxonomy for missing data mechanisms described for a special case of an outcome variable collected once at the end of the study

2.3.9.1 Missing At Random (MAR)

Consider a special case where data is missing only in the binary outcome variable Y, the outcome data are said to be *missing at random* (MAR) if the probability of being missing depends on the observed covariates, but is independent of the specific missing outcome values that should have been observed in principle. In mathematical terms this is expressed as follows:

$\Pr(D=1|Y,X)=\Pr(D=1|X)$ ……………………………………………………… …………….(2.14)

where Y , X are the complete outcome data, the covariates and D is a vector that indicates missingness such that $D=1$ if Y is missing and $D=0$ if Y is observed

Example

Consider a study in which efficacy (binary: 0=treatment succcess; 1=treatment failure), treatment (binary: 0=placebo; 1=Active treatment) and age (years) are the three variables that a researcher intends to observe on each participant. For now, consider only those study participants whose age is observed and equal to a particular value, say 10 years old and are on placebo. In practice efficacy may be missing for some of these participants aged 10 years and observed for others. MAR implies that among these study participants with observed age = 10 years and are on placebo, the chance of treatment success is the same among the cases for which efficacy has been ascertained as it is among the cases for

which efficacy is missing. Similarly, for study participants with observed age = 11 years, the chance of treatment success is the same among the cases for which efficacy is observed as it is among the cases for which efficacy is missing; however, the chance of treatment success for participants with observed age = 11 and are on placebo may be different from the chance of treatment success for those with observed age = 10 years and are on placebo. The same applies for participants with the specified observed ages above but who are on active treatment. This also applies to participants with observed age = 12 years, 13 years, …, etc conditional on their treatment status. That is, the missing efficacy should be regarded as a random sample of all the efficacy values within the observed age by treatment subgroups.

2.3.9.2 Missing Completely At Random (MCAR)

The outcome data are said to be *missing completely at random* (MCAR) if the probability of being missing does not depend on either the value of the outcome Y or the value of the covariate.

Mathematically this is expressed as follows:

$$\Pr(D=1 \mid Y, X) = \Pr(D=1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.15)$$

where Y , X are the complete outcome data, the covariates and D is a vector that indicates missingness such that $D=1$ if Y is missing and $D=0$ if Y is observed.

Example

Reconsider the example above, MCAR implies that the missing efficacy values are neither related to age, treatment on which someone is on nor to the efficacy itself (whether observed or not). Thus, with MCAR, the component of the data with missing efficacy is a random subset of the complete original sample of efficacy status – and equally, by definition, the observed sample is also a random sample of the original complete sample.

2.3.9.3 Missing Not At Random (MNAR)

The outcome data are said to be *missing not at random* (MNAR) if the probability that the outcome data are missing depends on both the observed outcome values and the unobserved outcome values. In mathematical terms this is expressed as follows:

$$\Pr(D=1|Y,\mathrm{X})=\Pr(D=1|\mathrm{Y}^{\mathrm{obs}},\mathrm{Y}^{\mathrm{mis}},\mathrm{X}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2.16)$$

2.3.10 Remarks on MAR, MCAR, and MNAR assumptions

There are important implication of the different missing data mechanisms: MAR, MCAR and MNAR (Rubin 1987, Allison 2001, Collins et al. 2001, Little 2002, Schafer and Graham 2002).

Statistical theory assumes that the data Y are randomly sampled from a distribution say $h(Y,\phi)$, where ϕ denotes unknown parameters governing the distribution of the data

Y. When the data Y is fully observed, $h(Y,\phi)$ describes both the sampling distribution for *Y* and the likelihood function for ϕ. Thus, when the data is fully observed for *Y*, the fact that $h(Y,\phi)$ may be regarded as a likelihood function for ϕ allows the application of Maximum Likelihood (ML) methods to obtain valid estimates of ϕ (Schafer and Graham 2002).

The situation is different in the presence of missing data. It is only under the MCAR assumption that the distribution of the observed data only denoted as $h(Y^{(obs)},\phi)$ can be regarded as both a correct sampling distribution of *Y* and a correct likelihood function for ϕ producing valid ML estimates of ϕ. In the presence of the missing Y, the distribution $h(Y^{(obs)},\phi)$ is not a correct sampling distribution of Y under MAR assumption. However, it is a correct likelihood function for ϕ under MAR (Rubin 1976, Schafer 1997, Little 2002, Schafer and Graham 2002). Thus when the data is MAR, the ML based estimation methods yield valid estimates of ϕ.

MCAR is the strongest assumption of the distribution of missingness – but is rarely satisfied in practice and is usually hard to justify. This assumption is usually met where the process that leads to the missing data has been made by design of the study (Kenward and Carpenter 2007).

Example

Consider an RCT to determine the efficacy of ant-malaria therapy by day 28 from the time participants took their first dose. The day 28 outcome would be assessed on all participants if there were no losses to follow up. It may also be possible that due to cost implications, a researcher may decide to follow up fewer participants up to day 42 and day 63 as secondary endpoints. A researcher can decide to take a random sample of 80% from the original sample. This means that some participants (20%) will have missing day 42 and day 63. If all the 80% of the original sample that has been sampled for further follow up will be successfully followed up without losses to follow up, then those that will have missing day 42 and day 63 observations will be MCAR.

MAR is less restrictive than MCAR and is generally plausible in practice (Collins et al. 2001, Kenward and Carpenter 2007). MAR assumption plays a critical role in the process of handling missing data because valid inferences can be made without regard to the missing data mechanism (Rubin 1976, Little 2002, Schafer and Graham 2002, Carpenter et al. 2007, Kenward and Carpenter 2007).

2.3.11 Missing data patterns

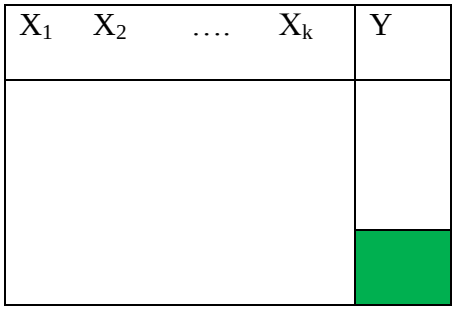
A missing data pattern is the way in which the observed and missing values are arranged in a data set. The missing data patterns that include: univariate pattern, monotone pattern, general pattern (Enders 2010). Schafer and Graham (2002), Van Buuren (2007) and Enders (2010) provide excellent descriptions and graphical presentations of the missing

data patterns. For the purposes of this thesis univariate pattern, monotone pattern and General pattern will be discussed.

2.3.11.1 Univariate missing data pattern

This is a data configuration such that data is fully observed for variables X_1, X_2,…, X_k but is missing for some participants for variable Y (Schafer and Graham 2002, van Buuren 2007, Enders 2010). This pattern may arise in randomized controlled trials where baseline variables are rigorously measured at baseline and form part of inclusion/exclusion criteria but may be missing for an outcome variable that is measured during follow-up. Figure 2.1 below gives a graphical presentation of a univariate pattern. In general this is not a common pattern in many study designs. Even in well designed randomized studies presence of some missing data in the baseline covariates is inevitable.

Figure 2.1: Illustration of a univariate missing data pattern based on (Schafer and Graham 2002, Enders 2010)

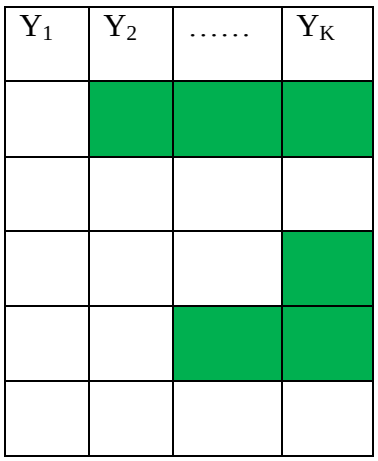


(The shaded area represents the missing Y data.)

2.3.11.2 Monotone missing data pattern

This pattern usually occurs in studies where participants are followed over time. Consider data Y_1, Y_2, ..., Y_k that are longitudinally collected at times t_1, t_2, t_3,….., t_k respectively. Monotone pattern means that if data is missing for Y_i then it will also be missing for Y_{i+1}…..Y_K. This missing data pattern is illustrated in figure 2.2 below.

Figure 2.2: Illustration of a monotone missing data pattern based on (Schafer and Graham 2002, Enders 2010).

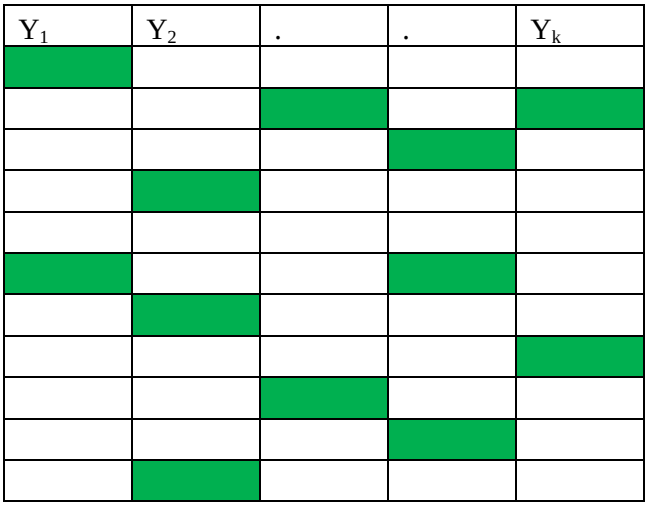


(The shaded area represents the missing data.)

2.3.11.3 General missing data pattern

Consider a dataset with several variables. In the general missing data pattern, data may be missing in any variable. The missing values are scattered all over the dataset. Although it is difficult to visually determine a clear pattern, it may still be a systematic pattern (Enders 2010). This is probably the most common pattern in practice. The pattern is illustrated in figure 2.3 below.

Figure 2.3: Illustration of a general missing data pattern *(the shaded area represent missing data)*



2.4 Common approaches for handling missing data

2.4.3 Unprincipled (adhoc) methods

The unprincipled statistical methods are methods for handling missing data which are not based on statistical models (Kenward and Carpenter 2007). In these methods the data are manipulated such that analyses proceed as if data was completely observed without paying attention to the process that is creating missing data (Kenward and Carpenter 2007). These methods often yield invalid inferences (Kenward and Carpenter 2007). Despite the existence of a variety of principled statistical methods of handling missing data, missing data is quite commonly handled using unprincipled methods.

There are a number of unprincipled methods and some common ones include: Complete Case analysis; Last Observation Carried forward; and Extreme case analysis.

2.4.4 Complete case analysis

Complete case analysis is the commonest approach to analysing incomplete data (Klebanoff and Cole 2008). This approach simply discards all cases with any missing values from the analyses. This approach has obvious advantages. Almost all standard statistical methods for analysis presume that all subjects have measurements on all variables included in the analyses (Allison 2001, Altman and Bland 2007) and therefore perform complete case analysis as a default analysis approach. In addition complete case analysis approach greatly simplifies the analytical process such that statistical analyses proceed as if there were no missing data at all. However, the use of this approach may have far reaching statistical consequences for the inferences. Despite its simplicity, complete case analysis approach is only valid under the assumption that missingness of data is not related to any variable in a dataset (MCAR scenario) (Little 2002, Molenberghs et al. 2004). However, the MCAR assumption may not be easily justified in practice, rendering the use of complete case analysis questionable in many cases. Even where MCAR assumption may be plausible, analyses from complete case analysis would suffer from loss of statistical power due to the reduced sample sizes. In the event that data are not MCAR, as is often the case, estimates from complete case analysis may be biased and would be inefficient, especially in multivariable analyses (Desai et al. 2011). In multivariable analyses data may be missing in many variables and a complete case analysis will discard a big proportion of the data thereby drastically reducing the sample size on which the analyses are based hence inefficient estimates will be obtained (Desai et al. 2011). In addition, complete case analysis approach is not consistent with the intention to treat principle which is the standard approach for analyzing data from

randomized controlled trials (Altman 2009). The intention to treat principle requires that all subjects that were randomized in a study are included in the analysis according to their randomization.

2.4.5 Last observation carried forward (LOCF)

LOCF is also one of the commonest solution for analyzing continuous outcome data with some missing observations from randomized studies which are longitudinal in design (Altman 2009). Missing data for subjects that dropout in longitudinal studies are replaced by the last observed measurement taken before the participant dropped out of the study. This method greatly simplifies analyses but is highly prone to producing biased estimates (Streiner 2008, Altman 2009). It assumes that from the time the last observation was taken, the value would have remained the same over time. This is not plausible in many cases (Shapiro 2001, Streiner 2008, Altman 2009). Observations are usually variable within an individual over time. The main advantage of this approach in randomized studies is that it allows application of the Intention to treat principle. However it should be noted that although the method is compatible with ITT principle, the estimates may be biased, making the inferences difficult to generalize (Lee et al. 1991, Shapiro 2001, Streiner 2008, Marshall et al. 2009).

Example

Suppose that in a longitudinal malaria therapy study, two anti-malaria treatments are administered to participants. Let measurements be taken on days: 0 (baseline), 1, 3, 7, 14

and 28. Let the outcome of interest be adequate clinical and parasitological failure. A subject dropping out of the study after day 3 measurement and before day 7 will have the day 3 measurement as the last observation. The is participant has a higher chance of having treatment failure on day 3 because of pharmacokinetic reasons and it may be difficult to justify an assumption that the outcome of this particular participant would have remained the same up to day 28. Furthermore, if the dropout is in the placebo/control treatment group the resulting estimates of treatment effect may be biased in favour of the active treatment when LOCF approach is used (Streiner 2008). In longitudinal studies it is very unlikely that the last measurement would have remained the same up to the end of follow up because correlation between repeated measurements tend to decrease with increasing time separation (Diggle et al. 2002, Hedeker and Gibbons 2006).

2.4.6 Extreme case (EC) analysis

This is another common approach for imputing missing binary outcome in randomized controlled studies (Lachin 1999, Higgins et al. 2008, Altman 2009). The missing values are replaced with either best-possible values or worst-possible values. For a continuous variable the largest and the smallest effect estimates that are consistent with the observed quantities are imputed to replace the missing values (Higgins et al. 2008). The estimates of treatment effect from the EC analyses are often biased (Lachin 1999). It is often difficult to justify that all subjects with missing outcome had the worst treatment or the best treatment effect. This approach is only useful for sensitivity analyses (Sterne et al. 2009).

2.4.7 Single imputation methods

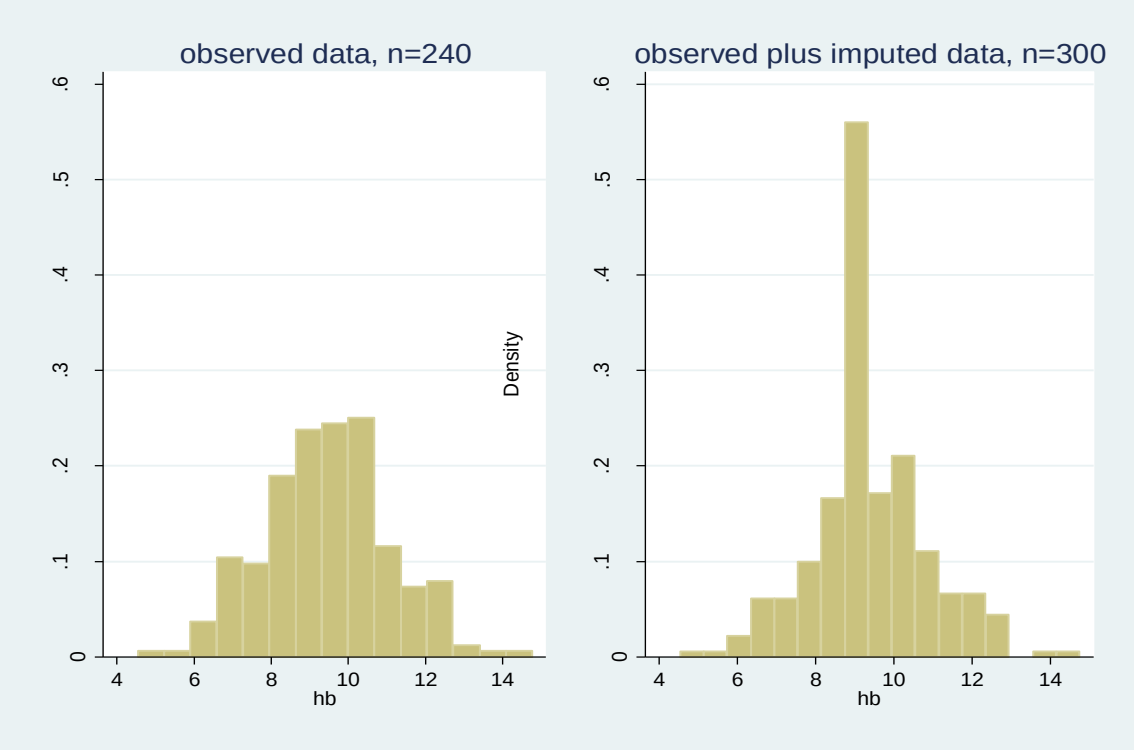
2.4.7.1 Mean imputation

In the mean imputation each missing value is replaced by the marginal mean. The main problem with this approach is that it alters the shape of the distribution of the imputed variable (figure 2.4). This method produces biased parameter estimates of location and in addition, it does not account for uncertainty due to imputed values (Enders 2010). Of course the obvious advantage is that it results in complete data which can then be analysed by any standard method of analysis.

Example

Figure 2.4 below illustrates how the distribution is distorted. A total of 300 hb values were simulated with mean=9.2g/dl. Missing values were randomly imposed in 20% of the observations. The missing values were then replaced by the mean of the observed values.

Figure 2.4: Histograms of the observed data and the complete marginal mean imputed data



2.4.7.2 Hot Deck imputation

This is a common imputation method in surveys. Suppose that data are measured on two variables X and Y such that variable X is complete, but variable Y has some missing observations. The hot deck imputation proceeds as follows: Firstly group subjects according to values of the complete variable X; then the missing Y observations in each group created in first step are replaced by the randomly sampled observed Y values in that subgroup.

The obvious advantage of this method is that it results in complete data that can be analysed by any standard statistical method. In addition it is consistent with the intention to treat principle. It will always result in a plausible range of results (Andridge and Little 2010). However the method lacks theoretical basis (Andridge and Little 2010). Just like other single imputation methods, this method results in small standard errors due to the fact that uncertainty in the imputed values is not taken into account in the substantive analyses because the imputed values are treated as if they were actually observed.

2.4.7.3 Regression imputation - Buck's method

This method was proposed by Bulk (1960). It obtains information for imputing missing values from other observed variables in the dataset (Buck 1960). It is also referred to as conditional mean imputation (Enders 2010). Let $Y, X_1, X_2, ..., X_p$ be the p+1 variables in a dataset such that Y values are missing for some participants and the variables $X_1, . . . , X_p$ are fully observed for all participants in the dataset. Firstly, the following regression model is fitted to the observed data as follows:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \quad \text{……………………………………………………………}(2.17)$$

The missing Y value for individual *i* is then imputed using estimates from the regression equation below:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_p X_{pi} \quad \text{…………………………………………………………...}(2.18)$$

The main advantage of this method is that it results in complete data. On the other hand, the main weakness of this approach is that correlations are inflated because all the

imputed values lie on the fitted regression line without random deviations from the line (Enders 2010). It is possible that some of the imputed values may lie outside a plausible range using this imputation method. The imputed values lack the variability that would have been present if the data had no missing values and this leads to biased parameter estimates of location of Y (Enders 2010).

The method provides consistent estimates under MCAR and MAR mechanism (Little 2002). The major problem with this approach is that variability is underestimated because the uncertainties about the imputed values are not taken into account.

2.4.7.4 Stochastic regression

This method is an extension of the regression imputation technique that aims at reducing correlations experienced in regression imputation procedure by adding a random error term to the conditional mean imputation equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_p X_{pi} + \varepsilon_i$$

……..…(2.19)

where $\varepsilon_i \sim N(0, \sigma^2)$ and σ^2 is usually replaced by S^2 in practice, the mean square error for the fitted regression model.

This method produces unbiased parameter estimates under MCAR and MAR assumptions, however, the resulting standard errors are small resulting in inflated type

one error (Enders 2010) because, in the substantive analyses, the imputed values are treated as if they were actually observed.

2.4.7.5 Propensity score method

The propensity score for a participant is the probability that the value of a variable is missing, conditional on the values of the other variables for that individual (Rosenbaum and Rubin 1983). The missing values of the variable are then imputed using observed values of that variable from other individuals for whom the variable has the same probability of being missing (Rosenbaum and Rubin 1983, Rosenbaum and Rubin 1984, Mattei 2009). This method is often used for dealing with a monotone missing data pattern.

Example

Let y=(Y_1, Y_2 , Y_3, Y_4) be the variables of interest such that Y_1 is observed for all cases while Y_2 , Y_3, Y_4 have some missing values such that if subjects have missing Y_2 will also have missing Y_3 and Y_4 ; and those with missing Y_3 will also have missing Y_4 (the monotone missing data pattern). The imputation is done sequentially such that first the observed values of Y_1 are used to impute the missing values for Y_2, then the values of Y_1 and Y_2 inclusive of the imputed values are used to impute the missing values for Y_3, and then the values of Y_1, Y_2 and Y_3 inclusive of the imputed values are used to impute the missing values for Y_4.

The propensity score method uses the following steps to impute values for each variable Y_i with missing values:

1. A missing data indicator variable D_j is created with the value 1 if Y_i is missing 0 otherwise where j=2,…..,k and k is number of variables of interest.

2. A logistic regression model for D_j is then fitted:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + ... + \beta_{j-1} Y_{j-1} \text{ ……..(2.20)}$$

3. The predicted probabilities of Y_j being missing are obtained for each observation using the logistic regression model fitted in step 2 above. This predicted probability is called a propensity score for each observation.

4. The observations are divided into a fixed number of groups based on these propensity scores.

5. A bootstrap imputation is then applied to each group created in step 4 to impute the missing Y_j values using the observed Y_j values with the same propensity score as the missing Y_j values. The bootstrap approach is carried out as follows: suppose that in a particular group, there are n_1 observed values of a variable Y_j and n_0 missing values. Then the boot strap imputation proceed as follows

a. Firstly a random sample of size n_1 taken from the observed values of Y_j sampling with replacement.

b. Next a random sample of size n_0 is drawn for the missing values from the n_1 observations of Y_j sampled in step a above.

Steps 1 through 5 above are repeated sequentially for each variable Y_j with missing values. The logistic regression model in step 2 above uses both the observed and imputed values of the variables $Y_1, Y_2 .... Y_{j-1}$ as explanatory variables.

The main advantage of this imputation model is that the imputed values are within the plausible range of the observed data. Of course in the case of randomized trials, it also allows the use of the intention to treat principle. The main disadvantage is that the imputed values are treated as if they had been actually observed which results in underestimated standard errors.

2.4.8 Principled methods

The Principled methods of handling missing data are the methods which are based on statistical models (Kenward and Carpenter 2007). The principled methods include Maximum-Likelihood based methods; Multiple imputation procedures and weighting methods.

2.4.8.1 Maximum likelihood based methods

The likelihood based approach for handling missing data proceeds as follows:

Let $h(Y,\phi)$ be a data model and $f(D|Y,\eta)$ be the corresponding missing data mechanism where the data model parameter ϕ and the missing mechanism parameter η are not related. Valid likelihood estimates of the data model parameter ϕ can be obtained based on the observed data only $h(Y^{(obs)},\phi)$ without making reference to the process that gives rise to missing data as long as the missingness mechanism is MAR (Little 2002) and that ϕ and η are not related (Rubin 1987, Little 2002). The joint distribution of the observed data and the missing data mechanism can be obtained as follows:

$$h(Y^{(obs)},D,\phi,\eta) = \int h(Y^{(obs)},Y^{(miss)},D,\phi,\eta)\mathrm{d}Y^{(miss)}$$

$$= \int h(Y^{(obs)},Y^{(miss)},\phi)\, f(D|Y^{(obs)},Y^{(miss)},\eta)\mathrm{d}Y^{(miss)} \quad \text{.....(2.21)}$$

When data is assumed to be MAR or MCAR (also called ignorable), *D* is independent of $Y^{(miss)}$

$$h(Y^{(obs)},D,\phi,\eta) = f(D/Y^{(obs)},\eta)\int h(Y^{(obs)},Y^{(miss)},\phi)dY^{(miss)}$$

$$= f(D/Y^{(obs)},\eta)h(Y^{(obs)},\phi) \quad \text{.....(2.22)}$$

and when the parameters for the data model ϕ and the missingness model η are not related, $h(\phi,\eta)=h(\phi)h(\eta)$ (Rubin 1976, Rubin 1987, Little 2002)

Thus, likelihood based inference on ϕ can be based on the observed data as long as data are either MAR or MCAR; and data model parameter ϕ and the missingness model parameter η are not related. For this reason, when data are either MAR or MCAR, the

missingness mechanisms are referred to as ignorable because the likelihood-based analysis can proceed in the estimation of data model parameters ϕ without making reference to the missingness mechanisms (Rubin 1976, Rubin 1987, Little 2002). The Expectation Maximization (EM) algorithm is often used to obtain valid parameter estimates from likelihood based analysis in the presence of missing data. The details of the EM algorithm are provided in the next subsection.

2.4.8.2 The EM algorithm

The EM algorithm was formalized and explained by Dempster (1977). It is an iterative procedure used for obtaining maximum likelihood estimates of parameters (Dempster et al. 1977). It is particularly useful and has wide applications in missing data problems.

1. The first step of an EM algorithm is an estimation of missing data to obtain a complete data set. i.e. draw

   $Y^{(miss)}$ from the predictive distribution of $Y^{(miss)}$ given the observed data $Y^{(obs)}$: $f(Y^{(miss)} \mid Y^{(obs)}, \phi^{(\ell)})$.

   Then repeat the following steps below:
2. Estimate model parameters using complete data set obtained in step 1 by averaging the complete data likelihood $L(\phi / Y)$ over $f(Y^{(miss)} \mid Y^{(obs)}, \phi^{(\ell)})$
3. The missing values are then re-estimated, based on parameter estimates obtained in step 2. Then go back to step2. This cycle is repeated several times and it stops when the estimates of parameters from the successive iterations remain unchanged.

2.4.8.3 Multiple imputation

Multiple Imputation is a statistical method for dealing with missing data that was devised by Rubin (1987). The MI process imputes ***p***>2 values for each missing value in a dataset based on information from other variables in the dataset. The imputed values that replace the missing values are repeatedly randomly drawn from the predictive distribution of $Y^{(miss)} \mid Y^{(obs)}$ (Rubin 1987). This essentially means that **p** different full datasets are created by the MI procedure. The MI approach ensures that the imputed values are not treated as if they had been actually observed when calculating standard errors. That is, the uncertainty about the true value that is missing is accounted for in the MI process when calculating standard errors using the observed and the imputed values. Both the between dataset imputation and within dataset imputation variability are taken into account. In the procedure, the overall estimate of the data model parameter say ϕ (where ϕ could be an estimate of a mean, proportion, regression coefficient etc) from the ***p*** datasets and the overall variance is obtained using Rubin's rules as illustrated below using notation similar to that of (Schafer and Graham 2002):

Let ϕ be the parameter of interest that needs to be estimated using the data (substantive) model and ϑ be the corresponding variance of the estimates in the absence of missing data. Further, let ***p*** different datasets be created by imputing ***p*** plausible values for each missing value. The estimate of ϕ is given by

$$\bar{\phi} = \frac{\sum_{i=1}^{p} \hat{\phi}^{(i)}}{p} \quad \text{.....(2.23)}$$

Where $\hat{\phi}^{(i)}$ are estimates from each of the **p** fully imputed datasets for i=1….p

The mean within imputation variance is given as:

$$\overline{\vartheta} = \frac{\sum_{i=1}^{p} \hat{\vartheta}^{(i)}}{p}, \quad \text{.....(2.24)}$$

where $\hat{\vartheta}^{(i)}$ are estimates of variance from each of the p fully imputed datasets for i=1….p and the estimate of between imputation variance ξ is obtained as follows:

$$\xi = \frac{\sum_{i=1}^{p} \left(\hat{\phi}^{(i)} - \overline{\phi}\right)^2}{p-1} \quad \text{.....(2.25)}$$

The overall estimate of the variance ψ is the combined within imputation and between imputation variances given as

$$\psi = \overline{\vartheta} + (1 + p^{-1})\xi \quad \text{.....(2.26)}$$

The confidence intervals and statistical test for the parameter ϕ are based on an approximation of a Student's t-statistic

$$\frac{(\overline{\phi} - \phi)}{\sqrt{\psi}} \sim t_V \quad \text{with} \quad v = (p-1)\left[1 + \frac{\overline{\vartheta}}{(1+p^{-1})\xi}\right]^2 \text{ degrees of freedom} \text{.....(2.27)}$$

For large samples, the 95 percent confidence interval for ϕ is estimated $\bar{\phi} + 1.96 * \sqrt{\psi}$ (van Buuren S. et al. 1999).

Under MAR assumption the estimates from multiple imputation are consistent, asymptotically efficient and asymptotically normal (Rubin 1987, Schafer 1997, Allison 2001, Little 2002). The advantage of MI over Maximum Likelihood method is that MI can be used with any type of data and any model type in most statistical software packages (Allison 2001).

Multiple imputations procedure can be implemented in the following steps:

1. **p** complete datasets are created by imputing missing values based on observed data but with some random variation introduced in the imputation process. Under MAR the imputations $y^{(1)}, y^{(2)}, ..., y^{(p)}$ are generated from the posterior predictive distribution of the missing given the observed values $f(y^{(miss)} \mid y^{(obs)})$. The resulting datasets will be slightly different from each other. Each of the of the p datasets can then be analysed using a standard applicable method e.g. binomial regression, logistic regression, poisson regression etc. This means that there will be p different estimates of the parameter of interest.
2. The estimates from the p datasets are combined into a single set of estimates. This procedure is done according to Rubin's rule as expressed above in equations 2.23-2.27.

There are many ways of achieving step 1. The methods are either from the Frequentist approaches or from the Bayesian paradigm. The common imputation models include: stochastic regression method, Propensity score method and Markov Chain Monte Carlo (MCMC)

The uncertainty due to imputation is reflected in the variability across the p estimates. Multiple imputation is one of the two principled methods (MI and ML) of handling missing data that is often much better than the older adhoc methods (Graham 2009). However, MI is simpler and more general to use than other principled methods because an analyst uses any standard statistical method that would be appropriate if the data were not missing (Schafer 1999). One major pitfall of MI approach is that its simplicity can easily make the user believe that the data are complete without regard to the missing data levels or mechanisms and this may be dangerous because the problem may be too trivial that use of MI may lead to significant biases between the complete data and imputed data estimates" (Dempster and Rubin 1983). Furthermore, it may be misleading to routinely consider MI as a method of choice under MAR assumption because the validity of this approach rests on the MAR assumption which is often difficult to test (Carpenter et al. 2007).

2.4.8.4 Imputation using chained equations

This imputation approach was devised by Van Buuren et al (1999). As described by carpenter and Kenward (2007), Let, *x, y,* be variables of interest both having some missing data. The imputation by chained equations approach firstly fill in the missing

values of *x* with randomly chosen observed values from *x*. The variable y is then regressed on x that contains both the observed and imputed values, and the missing y are filled in using the regression imputation. Similarly the observed x values are regressed on y that now contains both the observed and the imputed values. The values of x that were imputed by randomly choosing from the observed values in the first step are now replaced with the imputed values using the regression imputation. The procedure is repeated until convergence. This complete procedure constitutes one complete imputed dataset. Thus, the procedure is repeated m times to achieve m imputed datasets for the MI. This procedure is has some similarities with the Bayesian approach in the sense that the imputation of the values of a variable at sequence $y^{(l+1)}$ depends only on the values $y^{(l)}$. The Bayesian imputation is discussed below

2.4.8.5 Bayesian Multiple Imputation

The Bayesian MI models are detailed in Rubin (1987). Using notation similar to that of (Schafer 1999) , the Bayesian imputation proceeds as follows: Let *Y* be the only variable of interest with n observations such that $Y = (y_1, y_2, ..., y_j)$ be the j observed values of Y and $Y = (y_{j+1}, y_{j+2}, ..., y_n)$ are missing at random. The data may be partitioned into the observed and missing components.

Let $y_i \sim N(\phi, \sigma^2)$ for i=1.2,…,n and $\varsigma = (\phi, \sigma^2)$ ……..(2.28)

$P(\varsigma) \propto (\sigma^2)^{-1}$ under non informative prior ……(2.29)

The posterior distribution of ς for the observed data is given by:

$$\phi \mid \sigma^2, Y^{(obs)} \sim N(\bar{y}^{(obs)}, j^{-1}\sigma^2) \text{ ....}(2.30)$$

$$\sigma^2 \mid Y^{(obs)} \sim \frac{(j-1)S^{2(obs)}}{\chi^2_{(j-1)}} \text{ ....}(2.31) \text{ where}$$

$\bar{y}^{(obs)} = j^{-1}\sum_{i=1}^{j} y_i$, $S^{2(obs)} = (j-1)^{-1}\sum_{i=1}^{j}(y_i - \bar{y}^{(obs)})^2$ and χ^2_{j-1} is the usual chi-square statistic with j-1 degrees of freedom.

The missing $\mathrm{Y} = (y^{(\ell)}_{j+1}, y^{(\ell)}_{j+2}, ..., y^{(\ell)}_{n})$ values are then imputed as follows:

1. Firstly a random variance is simulated:

$$\sigma^{2(\ell)} \mid Y^{(obs)(\ell)} \sim \frac{(j-1)S^{2(obs)}}{\chi^2_{j-1}} \text{ ....}(2.32)$$

2. A random mean is then simulated:

$$\phi^{(\ell)} \mid \sigma^{2(\ell)}, Y^{(\ell)(obs)} \sim N(\bar{y}^{(obs)}, j^{-1}\sigma^{2(\ell)}) \text{ ....}(2.33)$$

3. The missing y values are then sampled independently:

$$y_i \sim N(\phi^{(\ell)}, \sigma^{2(\ell)}) \text{ for i=j+1......n ....}(2.34)$$

This procedure is repeated for $\ell = 2, ..., k$ where k is the number of proper imputations for $y^{(miss)}$ i.e. when the complete data estimator is the complete data maximum Likelihood Estimator (MLE).

This procedure is generalized as follows:

Let $Y = (Y^{(obs)}, Y^{(miss)}) \sim P(Y \mid \varsigma)$ where $Y^{(miss)}$ is ignorable (MAR or MCAR) and ς is unknown and has a prior distribution.

Since $f(Y^{(miss)} \mid Y^{(obs)}) = \int f(Y^{(miss)} \mid Y^{(obs)}, \varsigma) f(\varsigma \mid Y^{(obs)}) d\varsigma$, $y^{(miss)}$ is imputed as follows:

1. The first step is to simulate a random draw of ς from their observed –data posterior distribution

$$\varsigma^* \sim f(\varsigma \mid Y^{(obs)}) \quad \text{.....(2.35)}$$

2. The next step is to draw randomly the missing values of Y, Y^{miss} from their conditional predictive distribution

$$Y^{(miss)*} \sim f(Y^{(miss)} \mid Y^{(obs)}, \varsigma^*) \quad \text{.....(2.36)}$$

The Markov Chain Monte Carlo (MCMC) approach is increasingly being used for the Bayesian simulations for step 1(Schafer 1999). In very simplistic terms, the MCMC proceeds as follows :

1. Firstly an initial starting value for ς say $\varsigma^{(0)}$ is assigned.
2. The next step is to draw missing values of Y, Y^{miss} from the conditional predictive distribution $Y^{miss(\ell)} \sim f(Y^{(miss)} \mid Y^{(obs)}, \varsigma^{(\ell-1)})$ for $\ell = 1,2,....$
3. Then the next step is to sample ς from the posterior distribution of ς given $Y^{obs}, Y^{miss(\ell)}$:

$$\varsigma^{(\ell)} \sim f(\varsigma \mid Y^{(obs)}, Y^{mis(\ell)})$$

This creates a Markov Chain $(Y^{miss(\ell)}, \varsigma^{(\ell)})$ that converges to a stationary distribution $f(Y^{(miss)}, \varsigma \mid Y^{(obs)})$

The Gibbs sampler Algorithm is one of the common algorithms used in the MCMC. For variables $(Y_1, Y_2 ....., Y_n)$ in a dataset the Gibbs sampler proceeds by simulating from conditional distributions and proceeds as follow: Initialise all variables $(Y_1^{(0)}, Y_2^{(0)} .....Y_n^{(0)})$

1. The first step is to simulate a new $Y_1, (Y_1^{(1)})$ given the initialized values of $(Y_2^{(0)} .....Y_n^{(0)})$ i.e.

   $f(Y_1^{(1)} \mid Y_2^{(0)}, ..., Y_n^{(0)})$

2. The next step is to simulate a new $Y_2, (Y_2^{(1)})$ given the new, $(Y_1^{(1)})$ and the existing $(Y_3^{(0)} .....Y_n^{(0)})$

   $f(Y_2^{(1)} \mid Y_1^{(1)}, Y_3^{(0)}, ..., Y_n^{(0)})$

3. Then next simulate a new $Y_3, (Y_3^{(1)})$ given the new, $(Y_1^{(1)})$ and $(Y_2^{(1)})$ and the existing $(Y_4^{(0)} .....Y_n^0)$

   $f(Y_3^{(1)} \mid Y_1^{(1)}, Y_2^{(1)}, Y_4^{(0)}, ..., Y_n^{(0)})$ until all Y have been simulated.

This creates a single updated dataset. The updated datasets are sampled in similar steps to get new updates i.e. the next sample updates will be $f(Y_1^{(2)} \mid Y_2^{(1)}.....Y_n^{(1)})$, $f(Y_2^{(2)} \mid Y_1^{(2)}, Y_3^{(1)}.....Y_n^{(1)})$, $f(Y_3^{(2)} \mid Y_1^{(2)}, Y_2^{(2)}, Y_4^{(1)}.....Y_n^{(1)})$ for $(Y_1^{(2)})$, $(Y_2^{(2)})$ and $(Y_3^{(3)})$ respectively and the process continues until convergence. That is each new Y_j is updated conditional on the latest values of Y_j in the chain.

2.4.9 Weighting methods

Weighting procedures were initially developed for minimizing bias in surveys. In the presence of missing data there may be differential response between complete cases and those with missing observations. Weights are applied to the complete cases to minimize bias resulting from missing data. Inverse Probability weighting (IPW) is a common approach for weighting complete cases. The IPW method is often applied when estimates are obtained using Generalized Estimating equations (GEE) which are alternatives to maximum likelihood estimation especially in longitudinal studies. In GEE the weights are estimated as the inverse of the probability of being a complete case. This is a semi-parametric method and one needs to correctly specify the distribution of missingness in order to obtain consistent estimates of the parameter of interest (Robins et al. 1995, Rotnitzky and Robins 1997, Bang and Robins 2005).

When the missingness mechanism is not correctly specified, the inverse probability weighting tends to result in inconsistent estimates. A variation of the inverse probability weighting is known as doubly robust - inverse probability weighting (DR-IPW). The DR-

IPW produces consistent estimates when either the data model or the missingness model is correctly specified and not necessarily both (Scharfstein et al. 1999, Bang and Robins 2005). The unweighted estimates of parameters from GEE are not valid in the presence of missing data that is MAR (Little 2002, Kenward and Carpenter 2007). In the presence of missing data, the GEE parameter estimates are valid only when the missing data mechanism is MCAR (Kenward and Carpenter 2007). In order to obtain valid estimates under MAR when using GEE, weighting methods should be used (Little 2002, Carpenter and Kenward 2006, Carpenter et al. 2007).

2.4.10 Discussion of the methods of missing data

The methods of handling missing data are either principled or unprincipled. Unprincipled include Complete Case analysis, Extreme Case analysis, Last observation carried forward, and single imputation methods. Single imputation methods include: mean imputation, hot deck method, regression imputation method and stochastic regression method. Complete case analysis is probably the most common adhoc method that is applied in analyses. It is very simple to apply but the estimates of parameters from the CC method may be biased and are often in efficient. This method is only valid when data is missing completely at random, but even then, the estimates may be inefficient. Extreme case analysis is often applied in binary outcomes. In randomized trials this method has an advantage of being consistent with the intention to treat principle. However, just like CC, this approach often leads to biased estimates of parameters of interest. LOCF is a common technique for handling missing data in longitudinal studies. Although it is consistent with the intention to treat principle, the estimates from this method may be

biased. Both EC analysis and LOCF analysis approaches are often hard to justify. Single imputation methods allow use of all observations but they all underestimate standard errors because the uncertainty in the imputed values is not taken into account when estimating standard errors.

On the other hand a number of principled approaches exist which are either multiple imputation based, maximum likelihood based or weighting based. These methods lead to valid estimates of parameters when data is MAR. For example, under MAR assumption the estimates from multiple imputation are consistent, asymptotically efficient and asymptotically normal (Rubin 1987, Schafer 1997, Allison 2001, Little 2002). Unlike Maximum Likelihood method, MI can be used with any type of data and any model type in most conventional software (Allison 2001) because the EM algorithm. MI is more flexible and because of this it is often used in sensitivity analyses (Kenward and Carpenter 2007, Groenwold et al. 2011). On the other hand MI is not robust when the imputation model is misspecified. The same is true for the inverse probability weighting method but the doubly robust inverse probability weighting method offers a double protection to misspecification of either the imputation model or the data model (Scharfstein et al. 1999, Bang and Robins 2005, Carpenter and Kenward 2006, Machekano et al. 2008).

The critical issue in use of these methods is that it is often difficult to precisely test whether data is missing at random against missing not at random. It is therefore important to identify methods that are valid / robust in specific scenarios.

Chapter 3 : Methodology

3.0 Methodology

Two broad sets of simulation studies were performed in this project. The first set of simulations investigate the degree of model failure when modelling efficacy (risk) differences using the standard binomial regression model with an identity link function, and aimed at identifying alternative modelling approaches. The second set of simulations address the main objectives of this project that compares the performance of the Complete Case analysis and Multiple Imputation procedure over a range of efficacy scenarios for dealing with missing binary outcomes. All the studies were done using computer simulations. The parameters used to simulate data were based on the estimates of a malaria efficacy RCT dataset described in Chapter 1 above. The purpose of this was to simulate data that reflects real data as much as possible.

3.1 Choice of variables in the simulated data

Varying combinations of the following variables were included in the simulations described in this Chapter: group (the treatment that the participant was allocated to), weight, age, haemoglobin level (hb) and parasite density (parasitaemia). Group was included because the primary objective of the original study on which the simulations were based was to estimate the efficacy difference (effect size) between the two intervention groups. Age is often considered as a potential confounder in epidemiological studies so was considered to be an essential element of the substantive

model. Two other variables considered likely to be closely associated with outcome in malaria studies were haemoglobin and parasite density, so both were considered in simulations. Finally, as the dose of many anti-malarial treatments is adjusted for patient weight, weight was likely to be associated with outcome so was also included in the simulations. At the planning phase of the simulation exercises, it was expected that parasitaemia and weight would be included in imputation models primarily when missing outcomes could be considered to be "missing at random" so would provide important information in imputing the outcome variable.

3.1 Simulating datasets

In general datasets have been generated to mimic the real data that has been discussed in detail in Chapter 1 above. Below are the details of how data were simulated to mimic the real data.

3.1.1 Number of simulated datasets and sample sizes

A total of 5,000 data sets were simulated for each scenario that was examined. The scenarios that were investigated are described in the subsequent sections below. The sample size was 200 participants in each of the 5,000 simulated datasets in a scenario. The baseline covariate data was simulated for 200 participants. The two hundred participants were then randomized, in blocks of size 10, to two treatment groups: A and B in the ratio 1:1. One group represented a control treatment while the other was representing an intervention treatment arm.

3.2 Characteristics of the simulated dataset

The parameters used in the simulations were similar to those of an example RCT previously described in Chapter 1. The following baseline variables: age, weight (wt), haemoglobin (hb) level, parasitaemia count (para) were simulated for each treatment group. In order to simulate a multivariate normal distribution for the baseline covariates, the original age; wt and para data were first transformed into logarithmic scale (Marshall et al. 2010). The estimates of the means, variances and covariances were estimated based on the log transformed variables and are detailed in the next subsection below. The estimated parameters on log scale were used in data simulations. To maintain the skewness of these covariates that would reflect real data, the simulated log-normally distributed variables were transformed back to their original scales by taking an exponential function of each of these variables.

3.2.1 Parameter values for simulation of covariates

The matrices of parameters for simulating the baseline covariates were as follows:

$$\mathbf{X}=\begin{bmatrix} \log_e(Age) \\ hb \\ \log(Weight) \\ \log(Parasitaemia) \end{bmatrix} \quad \text{……..(3.1)}$$

$$\mu=\begin{pmatrix} 3.15 \\ 9.32 \\ 2.40 \\ 10.7 \end{pmatrix}$$

……(3.2)

$$\sigma = \begin{pmatrix} 0.42 \\ 1.66 \\ 0.18 \\ 1.50 \end{pmatrix} \quad \text{.....(3.3)}$$

$$\rho = \begin{pmatrix} 1.00 & 0.09 & 0.16 & 0.02 \\ 0.09 & 1.00 & 0.4 & 0.2 \\ 0.16 & 0.4 & 1.00 & 0.05 \\ 0.02 & 0.2 & 0.05 & 1.00 \end{pmatrix} \quad \text{.....(3.4)}$$

where:

X is a vector of the four covariates: logarithmic scale for - age, wt and para; and on original scale for haemoglobin (Hb);

μ is a vector of the mean values for log(age), Hb; log(weight) and log(parasitaemia) respectively.

σ is a vector of the standard deviation values for log(age), Hb; log(weight) and log(parasitaemia) respectively;

ρ is a matrix of the correlations between pairs of the baseline variables.

The matrix and vector values are derived from the historical data discussed in Chapter 1.

3.3 Randomization and the complete (full) dataset

The simulated observations with covariate data, generated as described in sections 3.1.3 above, were then randomized to two treatment groups: A and B in the ratio 1:1. The randomization was in blocks of size 10. After randomization of the simulated observations (that contain covariate data), the outcome variable was generated

(1=success, 0=failure) and treatment group (1=A, 0=B) as described in the next section 3.1.5 below.

3.4 Simulation of a binary outcome variable

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 a group i, for i=A, B. This resulted in a simulated binary outcome data with π_i success rate proportion (efficacy) and 1-π_i failure rate proportion in each group.

3.5 Investigating the Binomial regression model and alternative approaches for modeling efficacy (risk) differences

The primary aim of this thesis was to compare the performance of the CC analysis method and the multiple imputation method for analyzing binary outcome data to estimate an efficacy difference. The binomial regression model with an identity link function is the standard statistical modeling approach for this. When the binomial regression was fitted to the simulations to compare the performance of complete case (CC) and multiple imputation (MI) methods, it was observed that some of the models failed to converge. This caused great concern. Consequently, simulation studies were performed to investigate: the factors that are predictive of model failure; and to assess convergence and bias in the alternative approaches- the COPY method and the Cheung's modified OLS method.

3.6 Investigating the effects of proximity to boundary of prevalence levels and number of covariates on convergence of a binomial regression model

To investigate the factors that are associated with model failure, a simulation study was conducted as follows: the impact on the convergence of the binomial regression model of two factors, the proximity of one efficacy rate to the parameter boundaries (0% and 100%) and the number of covariates included in the binomial regression model, was examined.

Data was generated as described in section 3.1.1. Four efficacy scenarios were considered as follows: **60**% in group A and **70**% in group B; **60**% in group A and **80**% in group B; **60**% in group A and **85**% in group B; **60**% in group A and **90**% in group B. In these simulation studies, the efficacy rate in one group was being moved towards the boundary (100%)

For each efficacy scenario, one, two or three covariates were included in the substantive model. The rationale was to monitor the effect of moving one efficacy rate towards a boundary value and also the effect of increasing the number of covariates in a model.

3.7 The effect of correlations between covariates on model non-convergence

In this investigation, the correlations between covariates were removed. This was aimed at assessing whether the correlations between covariates have an impact on a risk difference model convergence. The correlations between any two covariates were set to

0.0. Exactly as for simulations based on the original correlations, for each efficacy scenario, one, two or three covariates were included in the substantive model.

3.8 Assessment criteria for factors associated with convergence

The percentage of the 5,000 datasets that converged were captured and summarised. The assessment was based on graphical methods. Line graphs were plotted to assess the presence of any trends in the in the percentage of the models that converge as one efficacy rates move towards a boundary parameter and also as the number of model covariates increases. The corresponding bias was also described using linear graphs.

3.9 The "COPY method" and the binomial regression model

The copy method was first proposed by Deddens and Petersen (2003) to address the problem of non-convergence when estimating risk ratios with the log-binomial model using Maximum Likelihood Estimation (MLE), which usually occurs when the risk ratio estimate is on the boundary of the parameter space (i.e. when either or both of the individual risk estimates is close to either 0% or 100%, so the ratio itself is either close to zero or heading off to infinity). As its name suggests, in this approach, multiple copies of the dataset are added to the original set, a small additional modification made (see below); when the binomial regression model is applied to this modified data set, the model converges and approximate maximum likelihood estimates of the risk ratio are obtained (Deddens and Petersen 2003, Deddens and Petersen 2008, Petersen and Deddens 2009).

In more precise statistical/mathematical terms, the copy method involves calculating MLEs using a log-binomial model on a new expanded version of the data set that contains K-1 copies of the original dataset plus one copy of the original dataset in which the values of the binary outcome variable are reversed (the 1's (successes) are all changed to 0's (fails) and the 0's (fails) are all changed to 1's (successes)). For a log-binomial model, if the total number of dataset copies, K, is finite, the iterative estimation solution is no longer on the boundary of the parameter space and is an MLE for the "copied" dataset (Petersen and Deddens 2009).

Petersen and Deddens (2008, 2009) state that, as K gets larger, the MLE estimate obtained from the "copied" dataset with a log-binomial model approaches the MLE estimate for the original dataset (i.e. is asymptotic), and they recommend that K should be at least 100 (although in their paper they used a value of K = 1,000). However, as the standard error estimates for the MLEs obtained with the copy method are based on K copies, they have to be multiplied by $\sqrt{(K)}$ to convert them to estimates for the original (single) dataset.

Mathematically, expanding the original data set in the manner required for the copy method is simply equivalent to creating a new data set consisting of one copy of the original data set having a weight of K-1 and one copy of the original data set with the outcome values reversed having a weight of one. Lumley (2006) states that use of the weights (K-1)/K and 1/K for the original outcome and the reversed outcome datasets

respectively eliminates the need to adjust the standard error (Lumley et al. 2006). The next section (3.1.12) describes how the COPY method was assessed.

3.10 The Assessment of convergence and bias of the COPY method

The covariate and outcome data was generated as described in section 3.1.1 above. The following efficacy rates were considered: 0.85 (85%) for group A and 0.60 (60%) for group B, a true absolute efficacy difference of 0.25 (25%); 0.98 (98%) for group A and 0.60 (60%) for group B, a true absolute efficacy difference of 0.38 (38%); and 0.98 (98%) for group A and 0.95 (95%) for group B, a true absolute efficacy difference of 0.03 (3%). The following copies were considered: 0 (no copy method), 10; 20; 50; 100; 500; 1,000; 1,500; 2,000; 3,000; 5,000; 10,000; 50,000 and 100,000.

3.11 The assessment criteria for the COPY method

Firstly, the percentage of simulated models that converged using the original dataset on its own was compared with the percentage of models that converged using the COPY method of the binomial regression model. The degree of bias in the MLEs of the true efficacy difference was then compared between the two analysis methods. Prior to considering the simulations studies for investigating convergence and bias of the copy method, a single original data set that failed to converge using the standard binomial model was assessed.

3.12 Cheung's modified OLS method

The Cheung's OLS, a seemingly potentially more reliable method than the Copy method was examined. The method fits the risk difference model using modified least-squares regression with a Huber-White robust standard error (Cheung 2007). Theoretically, this method should reduce the problem of model non-convergence that can occur when fitting a binomial regression model to obtain adjusted estimates of risk differences as it uses a different mathematical algorithm.

Cheung's modified method uses ordinary least squares (OLS) estimation together with Huber-White robust (H-W) robust standard errors. This method is reasonable if interest is confined to the estimation of risk differences, but is not suitable if there is interest in predicting probabilities for individual patients as estimated values outside the probability range 0 to 1 may be yielded. The method was considered for evaluation using simulations because the interest in this project was on estimating risk (efficacy) differences.

The method was examined using exactly the same efficacy scenarios as presented for the COPY method in the section (3.1.12) above. The method was assessed for bias and convergence percentage.

3.13 Comparison of methods of dealing with missing binary outcome data

3.13.1 The mechanisms for making data to be missing

Missing data was imposed on the binary outcome with specific missing rates as described in the missing data mechanism sections below. There was no missing data imposed on the covariates. This is so because in a randomized controlled setting, baseline variables are usually collected as part of the inclusion/exclusion criteria and as a result, they are rarely missing.

Three missingness data mechanisms were considered and these are: Missing at random (MAR*)*, missing completely at random (MCAR) and missing not at random (MNAR). The subsequent sections below describe how the missing mechanisms were generated.

3.13.2 Missing completely at random scenarios

To generate binary outcome data that is being missing completely at random, firstly a random variable is generated using a uniform [0,1] distribution. The random variable is then sorted and then the outcomes of the first p% of the participants are coded to have a missing outcome, P% takes the following values 5%, 15% and 30%. For example where p% is 5%, data is simulated such that in each of the 5000 datasets of size 200, there is a 5% missing binary outcome data (i.e. a total of 10 observations have missing outcome).

3.13.2.1 Missing at random scenarios

The following logistic regression model was used to generate data that is being missing at random:

$$\text{logit}(\pi) = 2 \times \text{group} + 0.277 \times \text{wt} \quad \ldots\ldots(3.5)$$

Thus, outcome data being missing was dependent on treatment group and weight of an individual as shown in the model above. There was no missing data imposed on the covariates as already explained.

Table 3.1 below presents a summary of data, regression (imputation) models and efficacy scenarios that were used in the simulations to compare the methods of dealing with missing data. The missing outcome data were imposed on each simulated dataset such that missing rates were set at 5%, 15% or 30% in each scenario being investigated (table 3.1) below. The specific missing rates were achieved using the missing model above combined with a uniform [0,1] distribution as detailed in stata do-files (appendix)

Table 3.1: Summary of data, regression (imputation) models and efficacy scenarios

	Description
Number of datasets	5,000
Sample size	200, (100 in each treatment group)
Variables in the dataset	Age, hb , wt* and para*, group and outcome
Variables in a risk difference regression model:	Continuous covariates: Age and hb . Factors: group(binary) Outcome: efficacy (binary)
Measure of effect :	Risk difference
Efficacy rates considered:	85 % in treatment A vs 60% in group B 98% in treatment A vs 60% in group B 98% in treatment A vs 95% in group B

*These variables are used only in imputation models

3.13.3 Rationale for choice of the efficacy rates

The efficacy rates were chosen in order to cover the three broad outcome scenarios that occur in practice. Firstly both rates were set to be reasonably away from the boundary. This scenario is possible in practice in malaria studies although not common but the scenario is likely to occur in many other studies that compare risk differences. An efficacy rate of 85 % in treatment A vs 60% in group B was considered for this scenario. In another scenario, one arm is set to have an efficacy rate that is close to the boundary while the other has efficacy rate away from the boundary value. This is a common

scenario in malaria efficacy studies where a new treatment with very high efficacy rate is compared with a standard treatment that has low efficacy rates probably due to resistance. An efficacy scenario of 98% in treatment A vs 60% in group B was considered to be a relevant scenario for this. Then in the third scenario that was considered, both efficacy rates were set to be close to a boundary value. This also commonly occurs in malaria studies especially in Phase II trials where both treatment regimens may have very high efficacy rates that are close to 100 (between 90% and 99%). An efficacy scenario of 98% in treatment A vs 95% in group B was deemed to be relevant for this. When modeling a risk difference, it is known that when efficacy rates are close to the boundary, the models tend to have convergence problems (Cheung 2007). The aim of considering these three scenarios is, therefore, to assess the performance of the multiple imputation approach and complete case analysis method in relation to the three general efficacy scenarios that occur in practice. It is anticipated that the methods of analysis may behave differently when handling missing data in these three different efficacy scenarios. Therefore the inclusion of the three possible efficacy scenarios when comparing the performance of the methods of handling missing data will help to generalize the findings.

3.13.4 Missing not at random scenarios

The Bernoulli distribution was used to generate data that is being missing not at random. The model was Bernoulli distribution was parameterized as Bernoulli (0.06) for a 5% missing data, Bernoulli (0.20) for 15% missing data and Bernoulli (0.40) for a 30% missing data. The model generated a 0 or a 1 for each observations and a further

condition was imposed based on this variable so that the missing data should depend on the value of the outcome. Then the outcomes were made to missing if they had a Bernoulli value of 1 and also if their outcome was 1 (a success outcome). Thus individuals with a success outcome (1s) were more likely to have their outcome missing. This would cause more individuals to have missing outcomes in the high efficacy group than the group with low efficacy which in turn results in would result in differential missing rates between the two groups.

Again the percentages of data being missing were set at 5%; 15% and 30% in each of the 5,000 datasets and in each scenario being investigated (table 3.2). The details are also found in stata do-files (appendix).

3.13.5 Model fitting

A risk difference regression model was fitted on full datasets as well as on incomplete datasets.

An adjusted mean estimate of a risk difference was obtained by fitting a risk difference model on the 5,000 full datasets. The resulting estimate was considered to be the true population adjusted risk difference.

The risk difference model was fitted using modified least-squares regression with a Huber-White robust standard error to obtain unbiased estimates of risk differences

(Cheung 2007). This method eliminates the problem of model non-convergence in software that sometimes results when fitting a binomial regression model to obtain adjusted estimates of risk differences (Cheung 2007). The method uses an ordinary least squares (OLS) estimation together with Huber-White robust (H-W) robust standard errors.

The simulated datasets with some missing binary outcome data were analysed using multiple imputation as well as complete case analysis. Age, haemoglobin and group were included in the data (substantive) models to obtain an adjusted risk difference between the control and treatment groups in the full data as well as in the versions with some missing outcome data. Several versions of multiple imputation models were performed under each efficacy scenario as well as under each missing rate scenario as detailed in table 3.2 below

Table 3.2: Summary of missing rates, missingness mechanisms, imputation models and model assessment criteria considered in the simulation studies

Assessment criteria	Bias, 95% CI and Coverage
Missing rates	An overall of 5%; 15%; and 30% in each dataset
Missingness mechanism	Missing At Random (MAR). Missingness dependent on weight and group such that those with high weight likely to have missing outcome in both groups. Those in the investigational treatment group were also simulated to have a higher chance of dropout. Missing Completely At Random (MCAR). Missingness of the outcome was unrelated to any variable whether observed or not. Missing Not At Random (MAR). Missingness on the outcome was dependent on the outcome as well as on group
Method comparison	Multiple imputation and Complete case analysis
Imputation models	1.logit(outcome)=$\beta_0+\beta_1$xwt+ $\beta_2$hb+$\beta_3$age+$\beta_4$para 2.logit(outcome)=$\beta_0+\beta_1$hb +$\beta_2$age+ $\beta_3$para 3.logit(outcome)=β0+β1group+β2hb+β3age+ β4para 4.logit(outcome)=β0+β1group+β2wt+β3hb+β4age+βpara

3.13.6 Assessment criteria

The main assessment criteria for comparison of missing data methods are the bias and coverage.

Let β be the parameter of interest. For the purposes of this thesis, this would be the population risk difference. In general this is not known in practice because it is difficult to study every member of the population of interest. The parameter β is usually estimated using a sample. The estimate may be denoted as $\hat{\beta}$ and this estimate may deviate from the true population value. Bias is the deviation of the parameter estimate $\ddot{\beta}$ from the true value β.

Mathematically bias denoted as δ is defined as follows(Burton et al. 2006):

$$\delta = \hat{\beta} - \beta \quad \text{.....(3.6)}$$

Confidence interval is given by $\beta \pm Z_{(1-\alpha/2)} \times SE(\hat{\beta}_i)$ .....(3.7)

Coverage is the proportion that the 100(1-α)% confidence intervals presented by $\beta \pm Z_{(1-\alpha/2)} \times SE(\hat{\beta}_i)$ include the parameter β wehre i=1…..N and Nis the number of simulations.

3.13.7 Software for simulations and analyses

Simulations and data analyses were performed using **Stata12**.1 software (**StataCorp**. 2009, Stata Statistical Software: Release 12, **StataCorp** LP,. College Station, TX, USA)

Chapter 4 : Alternative approaches to fitting binomial regression model

4.1 Chapter structure

This Chapter describes the problems encountered when using the binomial regression model with an identity link function to obtain adjusted risk/efficacy differences. Factors associated with the binomial regression model failure are investigated. Alternative approaches to the binomial regression models are considered using simulations and findings are presented.

4.2 Binomial regression model

The binomial regression model with an identity link function is the standard statistical method for analyzing binary outcome data to estimate a risk or efficacy difference. This regression technique is equivalent to fitting a generalized linear model from a binomial family with an identity link function.

When the binomial regression technique was employed during the first set of simulations to compare the performance of complete case (CC) and multiple imputation (MI) methods, it was found that about 5% of the 5,000 simulated datasets yielded non-convergent models when comparing efficacy rates of 85% in group A against 60% in group B. This was a cause for considerable concern. In a real trial situation, there is only one dataset. If a fitted substantive model does not converge for this particular dataset,

the data analyst needs to know the next steps that can be taken in order to evaluate the findings of the trial. This problem of non-convergence of binomial regression models has been reported previously (Wacholder 1986, Cheung 2007).

A search for methods that may be used to model risk/efficacy differences when the binomial model yields non-convergent results identified two potentially useful analysis policies: the copy method developed by Deddens (2003) and Cheung's Modified Ordinary Least Squares (OLS) method (2007). However, before evaluating these two methods, the factors that are associated with the binomial regression model failure when the identity link function is used were investigated.

4.2.1 Factors associated with the failure of a risk difference binomial regression model

The performance of the binomial regression method with an identity link function in the presence of a number of factors associated with a risk difference model failure was investigated as described in sections 4.2.1 and 4.2.2 below.

4.2.2 Effect of proximity to boundary of prevalence levels and number of covariates on convergence of a binomial regression model

Firstly, the influence on the convergence of the binomial regression model of two factors, the proximity of one or both prevalence levels to the parameter boundaries (0% and 100%) and the number of covariates included in the regression model, was examined. The findings are presented in Tables 4.1 and 4.2, and in Figures 4.1 and 4.2.

When using just a single predictor variable (covariate), the percentage of datasets that converged when a binomial regression model was fitted was high (96.8% to 99.7%) when the efficacy levels were 60% and 70% for the two groups respectively (i.e. were well away from 0% or 100%, the boundary of the parameter space). Stated the other way round, the percentage of datasets that failed to converge in this situation was very low (but not negligibly so).

Figure 4.1: Percentage convergence by efficacy rates and number of covariates in model

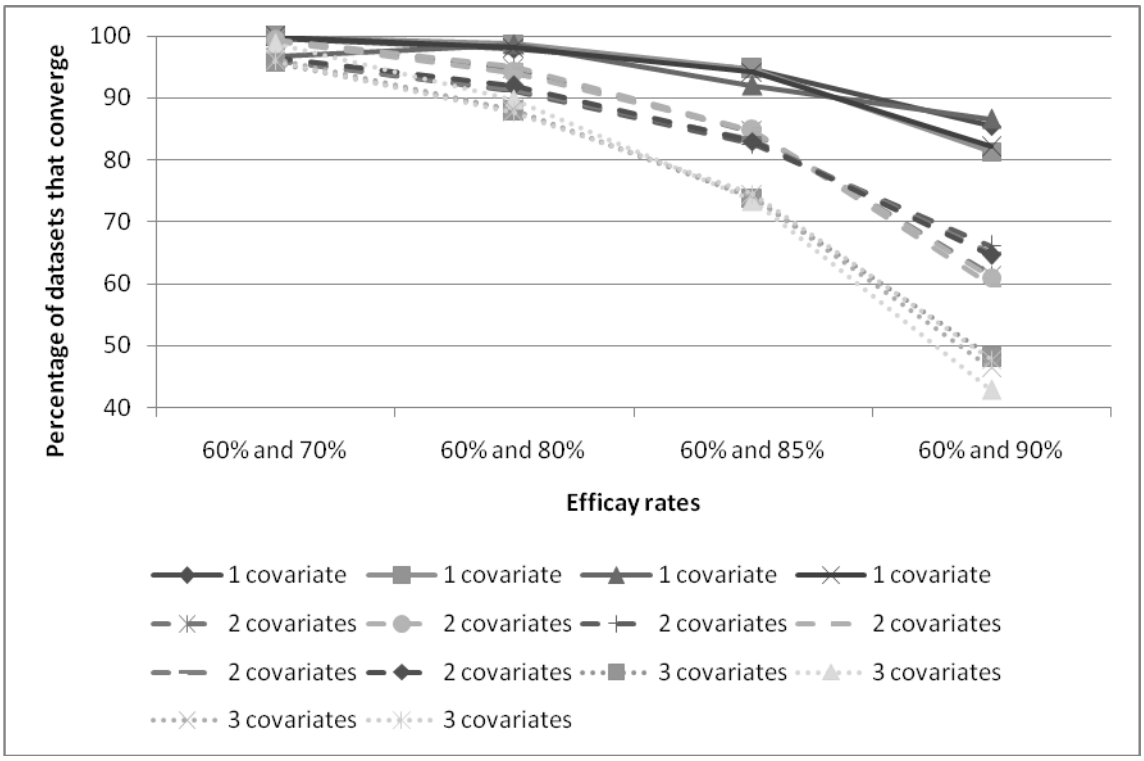


Table 4.1: Convergence rates by efficacy rate and number of covariates in model (averaged over 5000 simulated datasets)

Covariates:		Efficacy rates:			
Number	Names	60% vs. 70%	60% vs. 80%	60% vs. 85%	60% vs. 90%
1	age	99.7	97.9	94.6	85.5
1	hb	99.9	98.7	94.7	81.3
1	para	96.8	98.6	92.0	86.6
1	wt	99.9	98.1	94.2	82.1
2	hb, age	99.4	94.7	84.5	61.4
2	age, wt	99.3	94.1	84.7	61.0
2	age, para	96.0	91.1	82.6	66.1
2	hb, wt	99.4	95.1	84.7	59.8
2	hb, para	96.1	91.4	83.3	64.9
2	wt, para	96.4	92.0	83.0	64.6
3	age, hb, para	95.7	87.9	73.9	48.1
3	age, hb, wt	98.8	89.7	73.4	42.9
3	hb, wt, para	96.3	88.1	74.2	46.4
3	age, wt, para	95.9	87.6	74.5	47.7

Convergence rates remained high, ranging from 92.0% to 94.7%, even when the larger of the two efficacy levels were increased to 85%. However, when the larger efficacy level was increased to 90% (still some distance from the boundary), convergence levels were found to drop dramatically, to between 81.3% and 86.6%.

Adding additional covariates confounded rather than alleviated this problem. With the larger of the two efficacy levels set at 90%, convergence was obtained for between 59.8% and 66.1% of models involving two covariates, and for between only 42.9% and 48.1% of models involving three covariates.

In summary, non-convergence rates were found to increase as one or both of the efficacy rates moved towards a boundary value irrespective of the number of covariates included in the model. Indeed, it is interesting to note that the problem worsened with increasing numbers of covariates in the model and as efficacy moved towards a boundary value. That is, models with just one covariate had less convergence problems than models with two covariates, and in turn these had fewer convergence problems than models with three covariates. For all scenarios examined, convergence was poor when the efficacy rate in either group was 90%.

4.2.3 The effect of correlations between covariates on model non-convergence

In the previous section, the following levels of correlation were assumed between the covariates fitted in the different models: 0.4 for haemoglobin and weight; 0.1 for age and either weight or haemoglobin; 0.20 for age and weight; 0.02 for parasitaemia. These values reflected the levels found in the original dataset on which this simulation was based.

Figure 4.2: Effect of reducing correlation on convergence (Efficacy rates: 60% and 85%)

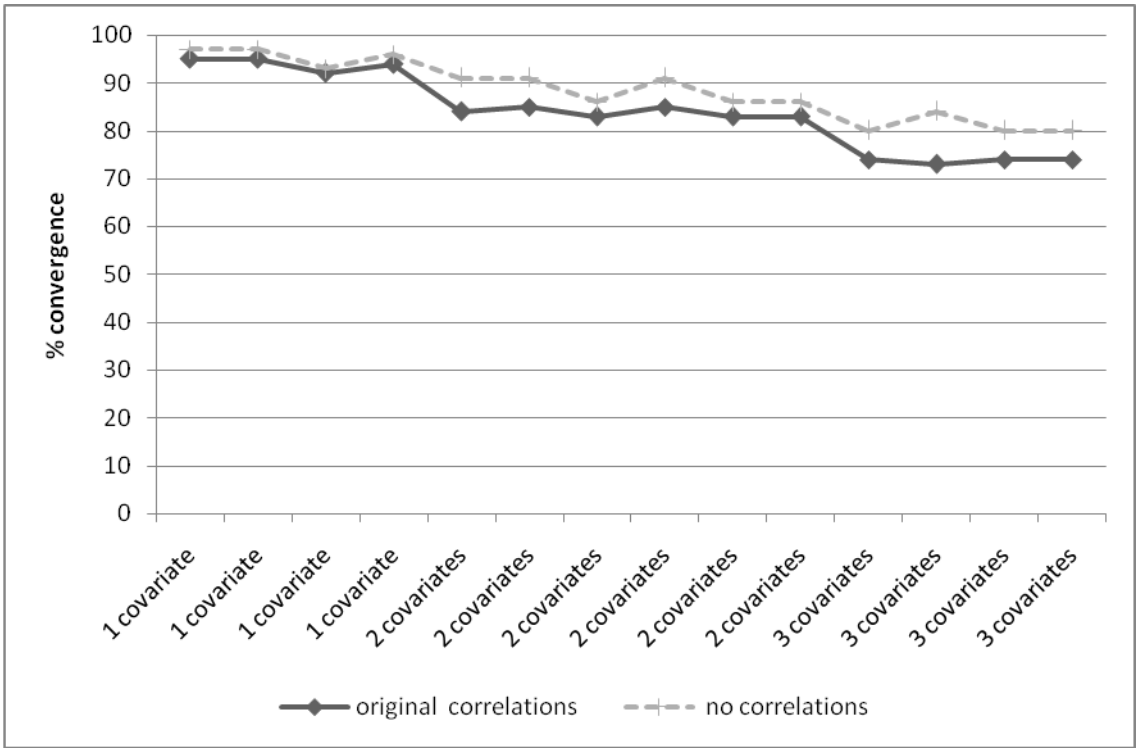


Table 4.2: Convergence rates in the presence and absence of correlation between covariates (averaged over 5000 simulated datasets): efficacy rates 60% and 85%

Number of covariates	Covariates	Original correlations	No correlations
1	age	94.6	97.2
1	hb	94.7	96.5
1	para	92.0	92.1
1	wt	94.2	96.3
2	hb, age	84.5	91.0
2	age, wt	84.7	91.1
2	age, para	82.6	85.7
2	hb, wt	84.7	91.1
2	hb, para	83.3	86.1
2	wt, para	83.0	85.8
3	age, hb, para	73.9	80.0
3	age, hb, wt	73.4	84.1
3	hb, wt, para	74.2	80.2
3	age, wt, para	74.5	80.4

The influence of this correlation was examined in more detail for efficacy rates in the two groups of 60% and 85% respectively, and the findings are summarised in Table 4.2 and Figure 4.2. In these simulations the correlations between any two covariates was set to zero. Similar findings were obtained for all other efficacy rate comparisons examined and so are not reported.

The percentage of datasets that converged improved when the correlations between the covariates were removed in all of the models considered. The improvement in convergence was most notable in models with higher number of covariates.

4.3 The "copy method" and the binomial regression model

The copy method was first proposed by Deddens and Petersen (2003) to address the problem of non-convergence when estimating risk ratios with the log-binomial model using Maximum Likelihood Estimation (MLE), which usually occurs when the risk ratio estimate is on the boundary of the parameter space (i.e. when either or both of the individual risk estimates is close to either 0% or 100%, so the ratio itself is either close to zero or heading off to infinity). As its name suggests, in this approach, multiple copies of the dataset are added to the original set, a small additional modification made (see below); when the binomial regression model is applied to this modified data set, the model converges and approximate maximum likelihood estimates of the risk ratio are obtained (Deddens and Petersen 2003, Deddens and Petersen 2008, Petersen and Deddens 2009).

In more precise statistical/mathematical terms, the copy method involves calculating MLEs using a log-binomial model on a new expanded version of the data set that contains K-1 copies of the original dataset plus one copy of the original dataset in which the values of the binary outcome variable are reversed (the 1's (successes) are all changed to 0's (fails) and the 0's (fails) are all changed to 1's (successes)). For a log-binomial model, if the total number of dataset copies, K, is finite, the iterative estimation solution is no longer on the boundary of the parameter space and is an MLE for the "copied" dataset (Petersen and Deddens 2009).

Petersen and Deddens (2008, 2009) state that, as K gets larger, the MLE estimate obtained from the "copied" dataset with a log-binomial model approaches the MLE estimate for the original dataset (i.e. is asymptotic), and they recommend that K should be at least 100 (although in their paper they used a value of K = 1,000). However, as the standard error estimates for the MLEs obtained with the copy method are based on K copies, they have to be multiplied by $\sqrt{(K)}$ to convert them to estimates for the original (single) dataset.

Mathematically, expanding the original data set in the manner required for the copy method is simply equivalent to creating a new data set consisting of one copy of the original data set having a weight of K-1 and one copy of the original data set with the outcome values reversed having a weight of one. Lumley (2006) states that use of the weights (K-1)/K and 1/K for the original outcome and the reversed outcome datasets respectively eliminates the need to adjust the standard error (Lumley et al. 2006).

Although the copy method is simple to apply and intuitively attractive, no published evidence could be found indicating whether the copy method can be extended for use with binomial regression models with the identity link function to obtain risk differences. In this project, therefore, the copy method was explored using simulation methods to assess whether its application can be extended to risk difference modeling when the original binomial model fails to converge.

4.4 Aims of the copy method assessment

- To assess whether the copy method resolves non-convergence problems for a binomial regression model with an identity link function.
- To assess whether the copy method produces unbiased estimates of risk differences when used with a binomial regression model.

4.4.1 Methodology for data simulations for the copy method assessment

The following matrices were used to simulate the covariate data based on the real data described in Chapter 3

$$X=\begin{pmatrix} \log_e(age) \\ Hb \end{pmatrix}$$

$$\mu = \begin{pmatrix} 3.15 \\ 9.32 \end{pmatrix}$$

$$\sigma = \begin{pmatrix} 0.42 \\ 1.66 \end{pmatrix}$$

$$\rho = \begin{pmatrix} 1 & 0.38 \\ 0.38 & 1 \end{pmatrix}$$

where: X is a matrix of the values for the two covariates (logarithmic) age and haemoglobin (Hb);

μ is a vector of the mean values for log(age) and Hb;

σ is a vector of the standard deviation values for log(age) and Hb respectively;

ρ is a matrix of the correlations between log(age) and Hb.

The outcome data were simulated using Bernoulli distributions with an efficacy rate of 0.85 (85%) for group A and 0.60 (60%) for group B, a true absolute efficacy difference of 0.25 (25%). The procedure was repeated for efficacy rate of 0.98 (98%) for group A and 0.60 (60%) for group B; and 0.98 (98%) for group A and 0.95 (95%) for group B.

4.4.2 The assessment criteria

Firstly, the percentage of simulated models that converged using the original dataset on its own was compared with the percentage of models that converged using the copy method. The degree of bias in the MLEs of the true efficacy difference was then compared between the two analysis methods. Prior to considering the simulations studies for investigating convergence and bias of the copy method, a single original data set that failed to converge using the standard binomial model was assessed.

4.4.3 Copy method for a single original data set

A single data set that failed to converge with the standard binomial regression method was investigated whether copy method would make the model converge. Table 4.3 below summarises the number of copies and whether the model converged or not.

Table 4.3: Summary of convergence using copy method for a single dataset

number of copies	RD	SE(RD)	status
0	*N/A*	*N/A*	*did not converge*
5	0.108	0.056	converged
6	0.114	0.055	converged
7	0.119	0.054	converged
8	0.122	0.054	converged
9	0.125	0.053	converged
10	0.127	0.053	converged
11	0.128	0.052	converged
12	0.130	0.052	converged
13	0.131	0.051	converged
14+	*N/A*	*N/A*	*did not converge*

For small number of copies (1 through 13 of original copies of outcome plus one copy of the original dataset with outcomes reversed), with three covariates Hb and age and weight, the analyses of the historical data converged. However models stopped converging when the number of copies of the original dataset exceeded 13. No statistical output was produced in Stata. However Stata reported "convergence not achieved". The number of iterations was then increased from the default 16000 to several thousands of iterations but no convergence was achieved. Tables 4.3 above summarise how the treatment effect size was changing with an additional copy of the reversed dataset (5-13 copies). The challenge was to know whether the treatment effect was moving towards the expected effect as the number of original copies increased. Unfortunately, the correct answer for "the expected effect" is unknown. Therefore one cannot be certain

whether the estimate from the 13 copies yield unbiased estimate of the risk differences and that the 13 copies are enough. A simulation study was performed to investigate the convergence problem and bias using a known treatment effect. The simulated data structure is described in section 4.3.2 above.

4.4.4 Bias and convergence rate trends for the COPY method simulations

4.4.4.1 Efficacy rates 85% vs. 60%

When 5000 simulated datasets were analysed using a binomial regression model with an identity link function *without* using the copy method, 4.9% (just under 5%) failed to converge (Figure 4.3, Table 4.4). In those models for which convergence was achieved, the mean efficacy difference estimate was 0.240 (s.e. 0.003), a (negative) bias of -0.010 (4.0%).

When the outcomes in one copy of the original dataset were reversed and then appended to increasing numbers of copies of the original dataset, the percentage of non-convergent models initially decreased. However, when the number of copies was extended beyond 10, the percentage of non-convergent models started to increase. At between 100 and 150 copies, the non-convergence rate reached the level observed when no copies were used, but then continued to increase as the number of copies rose further. Although the rate of increase diminished, the percentage of non-convergent model was observed to be still rising even when the number of copies used reached 100,000.

Figure 4.3: Percentage convergence and (absolute) bias for increasing numbers of copies (85% vs. 60%)

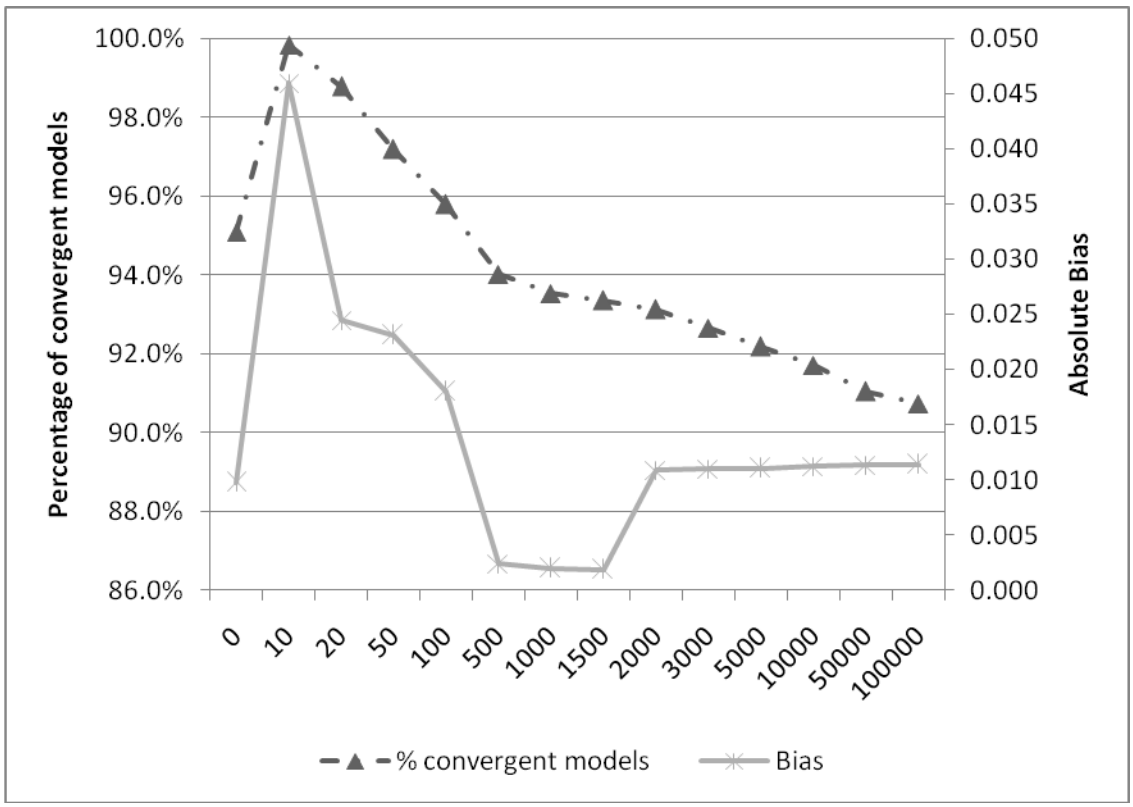


Table 4.4: Percentage convergence and bias for increasing numbers of copies (averaged over 5000 simulated datasets): 85% vs. 60% (RD = 0.250)

Number of copies	Converged		Risk difference (RD)		95% CI for RD		Bias
	n	(%)	Estimate	SE	LL	UL	
0	4755	(95.1)	0.240	0.003	0.235	0.245	-0.010
10	4992	(99.8)	0.204	0.001	0.202	0.206	-0.046
20	4939	(98.8)	0.226	0.001	0.224	0.227	-0.024
50	4860	(97.2)	0.227	<0.001	0.226	0.228	-0.023
100	4790	(95.8)	0.232	<0.001	0.231	0.233	-0.018
500	4701	(94.0)	0.248	<0.001	0.247	0.248	-0.002
1000	4677	(93.5)	0.248	<0.001	0.248	0.248	-0.002
1500	4668	(93.4)	0.248	<0.001	0.248	0.248	-0.002
2000	4656	(93.1)	0.239	<0.001	0.239	0.239	-0.011
3000	4633	(92.7)	0.239	<0.001	0.239	0.239	-0.011
5000	4609	(92.2)	0.239	<0.001	0.239	0.239	-0.011
10000	4586	(91.7)	0.239	<0.001	0.239	0.239	-0.011
50000	4552	(91.0)	0.239	<0.001	0.239	0.239	-0.011
100000	4537	(90.7)	0.239	<0.001	0.239	0.239	-0.011

CI: confidence interval LL: lower limit UL: upper limit

It was found that the datasets that failed to converge when analysed conventionally were predominantly the same as those that failed to converge using the copy method with

more than 100 copies. In addition, some datasets that converged when analysed conventionally became non-convergent using the copy method.

The degree of (negative) bias in the MLE estimates increased initially from -0.010 (4.0%) when no copies were used to -0.024 (9.6%) when 10 copies were used (the minus sign here indicating that the MLEs under-estimated the true efficacy difference). Bias levels then steadily decreased to just -0.002 (0.8%) when 500 to 1500 copies were used. With more than 1500 copies, bias then increased again, reaching a plateau level of -0.011 (4.4%) at around 2000 copies.

Ironically, and perhaps surprisingly, therefore, with the binomial model, the number of copies required to minimize the number of non-convergence models was found to coincide with the number of copies giving the most biased estimates of the true efficacy difference.

4.4.5 Copy method - bias and % convergence trends (95% vs. 90% efficacy rates)

When 5000 simulated datasets were analysed using a binomial regression model with an identity link function *without* using the copy method, 29% (about 30%) failed to converge (Figure 4.4, Table 4.5). In those models for which convergence was achieved, the mean efficacy difference estimate was 0.037 (s.e. 0.001), a (negative) bias of -0.013 (35%).

Figure 4.4: Percentage convergence and (absolute) bias for increasing numbers of copies (95% vs. 90%)

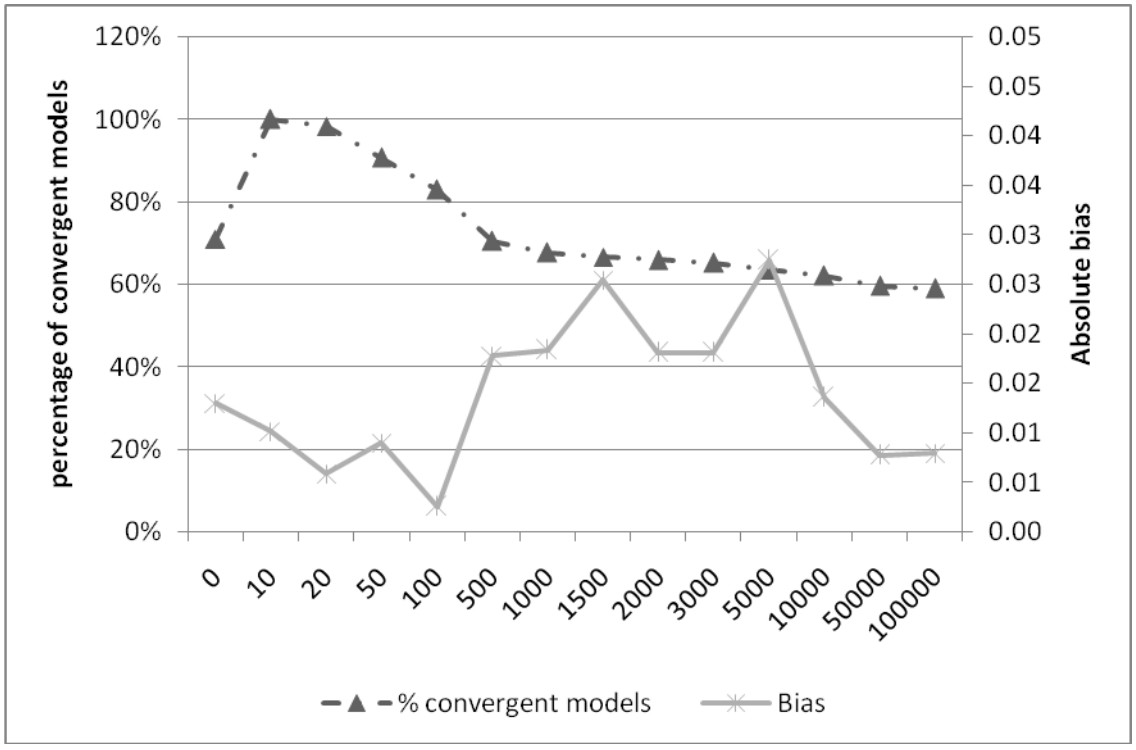


Table 4.5: Percentage convergence and bias for increasing numbers of copies (averaged over 5000 simulated datasets) : 90% vs. 95% (RD = 0.050)

Number of copies	Converged		Risk difference (RD)		95% CI for RD		Bias
	n	(%)	Estimate	SE	LL	UL	
0	3550	(71.0)	0.037	0.001	0.034	0.040	-0.013
10	4998	(100.)	0.040	0.001	0.038	0.041	-0.010
20	4917	(98.3)	0.044	<0.001	0.043	0.045	-0.006
50	4536	(90.7)	0.059	<0.001	0.058	0.059	+0.009
100	4143	(82.9)	0.047	<0.001	0.047	0.048	-0.003
500	3521	(70.4)	0.032	<0.001	0.032	0.032	-0.018
1000	3385	(67.7)	0.032	<0.001	0.032	0.032	-0.018
1500	3331	(66.6)	0.025	<0.001	0.024	0.025	-0.025
2000	3295	(65.9)	0.032	<0.001	0.032	0.032	-0.018
3000	3258	(65.2)	0.032	<0.001	0.008	0.008	-0.018
5000	3178	(63.6)	0.023	<0.001	0.023	0.023	-0.027
10000	3102	(62.0)	0.036	<0.001	0.036	0.036	-0.014
50000	2981	(59.6)	0.042	<0.001	0.042	0.042	-0.008
100000	2948	(59.0)	0.042	<0.001	0.042	0.042	-0.008

CI: confidence interval LL: lower limit UL: upper limit

Similar to the 85% against 60% efficacy rates, when the outcomes in one copy of the original dataset were reversed and then appended to increasing numbers of copies of the

original dataset, the percentage of non-convergent models initially decreased, reaching a zero non-convergence (100% convergence) rate with 10 copies. However, when the number of copies was extended beyond 10, the percentage of non-convergent models started to increase; again, at between 100 and below 500 copies, the non-convergence rate reached the level observed when no copies were used, but then continued to increase as the number of copies rose further. Although the rate of rise was gradual, the percentage of non-convergent model was observed to be still increasing even when the number of copies used reached 100,000, at which point the proportion of models not converging was approximately 60%.

As for the 85% vs. 60% comparison, the same datasets that failed to converge when analysed using the conventional approach failed to converge using the copy method with more than 100 copies – but in addition some datasets that converged when analysed conventionally became non-convergent using the copy method.

The degree of bias in the MLE estimates decreased initially from -0.013 (35%) when no copies were used to -0.010 (27%) when 10 copies were used. Bias steadily reduced to -0.003 (8%) with 100 copies of the original dataset (becoming positive briefly at around 50 copies). When more than 100 copies were used, the (negative) bias then increased again, fluctuating between -0.008 (21%) and -0.027 (73%).

Ironically again, with the binomial model, the number of copies required to achieve an absence of non-convergent models was found to coincide with the number of copies giving the most biased estimates of the true efficacy difference (a risk difference of 4% instead of the expected value of 5%).

4.4.6 Copy method - bias and % convergence trends (98% vs. 95% efficacy rates)

When both efficacy rates were set very close to boundary values, 60.3% of models failed to converge using a binomial regression model with an identity link function *without* using the copy method (Figure 4.5, Table 4.6); this is a large and wholly unmanageable convergence problem in practice. In those models for which convergence was achieved, the mean efficacy difference estimate was 0.036 (s.e. 0.001), a (negative) bias of -0.006 (17%).

Similar to the efficacy rates considered in the previous sections, when the outcomes in one copy of the original dataset were reversed and then appended to increasing numbers of copies of the original dataset, the percentage of non-convergent models initially decreased; again, a zero non-convergence rate (100% convergence rate) was observed with 10 copies of the original data set. However, when the number of copies was extended beyond 10, the percentage of non-convergent models started to increase. Again, at about 3000 copies, the non-convergence rate reached the level observed when no copies were used (a rate of about 40%), but then continued to increase as the number of copies rose further. The percentage of non-convergent model was observed to be

increasing even when the number of copies used reached 100,000, having at that stage almost reached 75% (i.e. barely one quarter of models were converging). Again, the datasets that did not converge when the copy method was not used also failed to converge using the copy method with more than 100 copies; also some datasets that converged when analysed conventionally became non-convergent using the copy method.

Figure 4.5: Percentage convergence and (absolute) bias for increasing numbers of copies (98% vs. 95%)

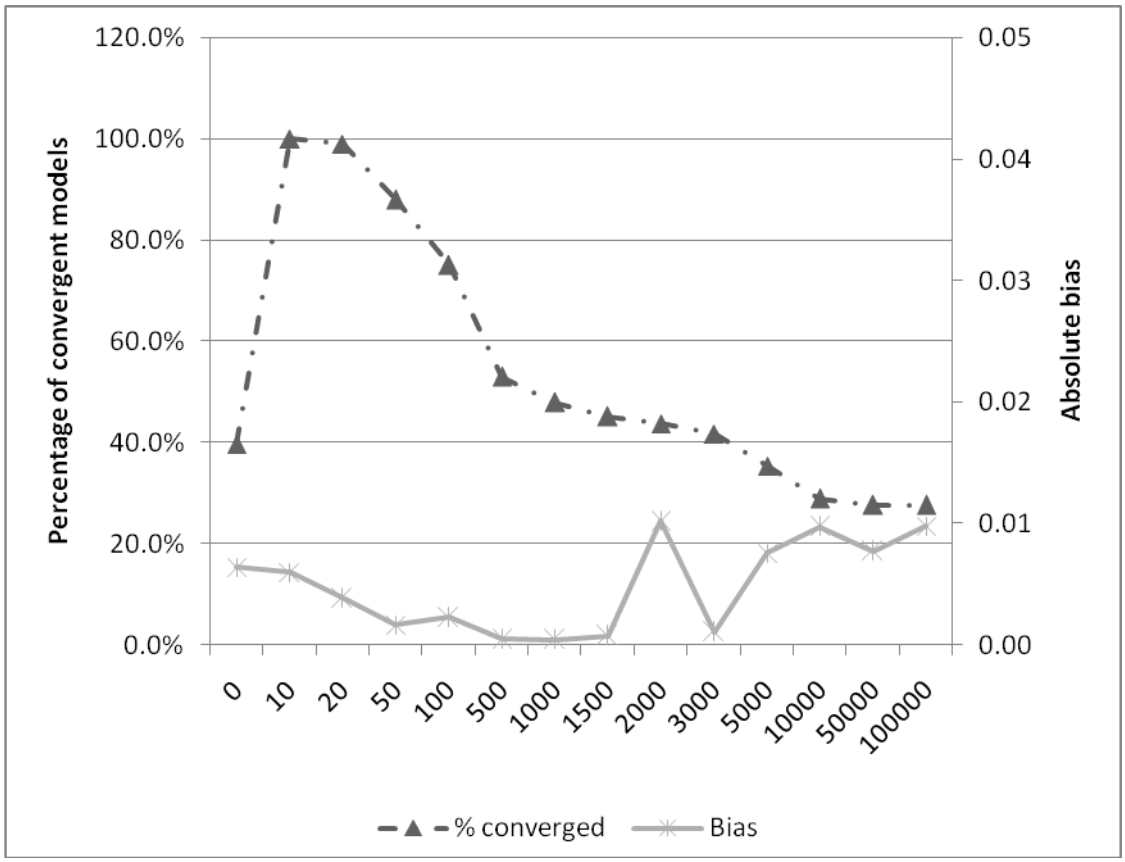


Table 4.6: Percentage convergence and bias for increasing numbers of copies (averaged over 5000 simulated datasets): 98% vs. 95% (RD = 0.030)

Number of copies	Converged	Risk difference (RD)		95% CI for RD		Bias
	n (%)	Estimate	SE	LL	UL	
0	1987 (39.7)	0.036	0.001	0.035	0.038	+0.006
10	5000 (100)	0.024	0.001	0.023	0.025	-0.006
20	4950 (99.0)	0.026	<0.001	0.025	0.027	-0.004
50	4403 (88.1)	0.028	<0.001	0.028	0.029	-0.002
100	3756 (75.1)	0.032	<0.001	0.032	0.032	+0.002
500	2649 (53.0)	0.030	<0.001	0.029	0.030	0.000
1000	2397 (47.9)	0.030	<0.001	0.029	0.030	0.000
1500	2255 (45.1)	0.029	<0.001	0.029	0.029	-0.001
2000	2188 (43.8)	0.040	<0.001	0.040	0.040	+0.010
3000	2089 (41.8)	0.029	<0.001	0.029	0.029	-0.001
5000	1977 (35.4)	0.038	<0.001	0.038	0.038	+0.008
10000	1772 (28.9)	0.040	<0.001	0.040	0.040	+0.010
50000	1446 (27.6)	0.038	<0.001	0.038	0.038	+0.008
100000	1380 (27.6)	0.040	<0.001	0.040	0.040	+0.010

CI: confidence interval LL: lower limit UL: upper limit

The degree of bias in the MLE estimates decreased initially from (an under-estimate of) -0.006 (20%), through zero bias at around 5 copies, and then to (an over-estimate of) +0.006 (20%) when 10 copies were used. Bias levels then slowly decreased, eventually fluctuating around zero and finally plateauing at zero when between 500 and 1000 copies were used. Bias then increased steadily, reaching a final plateau of around +0.010 (33.3%) from 5000 copies onwards.

The number of copies required to achieve zero non-convergence was again found to coincide with the number of copies giving highly biased estimates of the true efficacy difference (2.4% instead of the expected 3%).

In summary, irrespective of the sizes of the two efficacy rates being compared, the maximum percentage of convergent models was achieved using 10 copies. In general terms, when the number of copies was extended beyond 10, the percentage of non-convergent models increased steadily, continuing to do so even when the number of copies used reached 100,000. The number of copies required to achieve minimum non-convergence models coincided with the number of copies giving high biased estimates of the true efficacy difference.

These simulations indicate strongly that the copy method probably has no place in modeling risk differences using a binomial regression model, as in general both convergence levels and bias were found to be unacceptably high.

4.4.7 Cheung's Modified Ordinary Least Squares (OLS) method

As a potentially more reliable method than the Copy method, the results obtained by fitting the risk difference model using modified least-squares regression with a Huber-White robust standard error (Cheung 2007) were examined. Theoretically, this method should reduce the problem of model non-convergence that can occur when fitting a

binomial regression model to obtain adjusted estimates of risk differences as it uses a different mathematical algorithm.

Cheung's modified method uses ordinary least squares (OLS) estimation together with Huber-White robust (H-W) robust standard errors. This method is reasonable if interest is confined to the estimation of risk differences, but is not suitable if there is interest in predicting probabilities for individual patients as estimated values outside the probability range 0 to 1 may be yielded. The method was considered for evaluation using simulations because the interest in this project was on estimating risk (efficacy) differences. Table 4.7 presents the results for the Cheung's OLS method.

4.7: Percentage convergence and bias for Cheung's method

Efficacy rates	**Converged N(%)**	**RD (SE)**	**95% CI***		**Coverage**	**Bias**
			LL	**UL**		
85% vs 60%	5000 (100)	0.250 (0.061)	0.130	0.369	0.950	0.000
98% vs 60%	5000 (100)	0.380 (0.051)	0.279	0.479	0.946	0.000
98% vs 95 %	5000 (100)	0.030 (0.026)	-0.021	0.079	0.950	0.000

As expected of an OLS regression technique, Cheung's modified OLS method yielded 100% convergence rates and unbiased estimates of risk difference for all of the efficacy scenarios that were considered.

4.4.8 Conclusion

Cheung's modified OLS method was thus adopted for use in all of the simulations reported in the next Chapter to investigate methods for handling missing binary outcomes in a randomized controlled trial. This method avoids any problems of model non-convergence when estimating risk differences even when several covariates are included in the regression model, making it useful for controlling for potential confounders and also for identifying independent predictors of outcome when modeling risk differences. In addition, the method yields unbiased estimates of risk differences with robust standard errors, thus offering clear statistical advantages over the use of the binomial regression method with the Copy method.

When used with an identity link function to estimate either unadjusted or adjusted risk differences, the binomial regression model is susceptible to model non-convergence, particularly if one or both of the efficacy rates is close to a boundary value (i.e. is close to 0% or 100%). Increasing the number of covariates model often merely aggravates the problem, rendering the method inappropriate for adjusting for potential confounders. High correlations between model covariates also intensify the non-convergence problem.

The Copy method has been found to be effective in solving non-convergence problems in log-binomial models used to estimate risk ratios – but this research shows clearly that

the method is not appropriate for modeling risk differences. In all of the scenarios considered, 100% convergence was often achieved when around 10 copies were used – but this was found to be the number of copies at which the estimates for the risk difference were most biased. Increasing the number of copies beyond 10 simply increased the likelihood of non-convergence.

Interestingly, datasets that did not converge with the original binomial model also failed to converge with the Copy method, particularly when large numbers of copies were made. In addition some of the datasets that were convergent with the original binomial regression model become non-convergent with the Copy method. The possible reason for this is that the Copy method creates very large datasets. In general, very large and very small datasets are both susceptible to non-convergence problems (SAS Technical Support 2009).

In addition, bias patterns were found to be very irregular with increasing number of copies, a finding that needs further exploration. This was considered to be outside the limitations of this dissertation, the main aim was of which is to compare methods for dealing with missing binary outcome data. The non-convergence problems reviewed in this Chapter was an unexpected finding along the way and was pursued only as far as was necessary to ensure that the evaluation of missing data method comparisons could proceed smoothly.

Cheung's modified OLS method was thus adopted for the estimation of risk differences and for use in the simulation-based research reported in the next Chapter to compare methods for handling missing binary outcome data. This method was found to have excellent convergence properties, and produced unbiased estimates of both unadjusted and adjusted risk differences.

If convergence problem do not occur, however, the binomial model has one potential advantage in that it provides exact confidence intervals and so may be preferred. In Cheung's method, valid model-based estimates of standard errors (and hence of confidence intervals) are produced using Huber-White robust formulae, but as these are not based directly on the binomial distribution, they produces symmetrical 95% confidence intervals for estimates of both the individual group risk levels and the risk difference. Using the binomial distribution, these intervals are asymmetrical.

For estimating risk differences (which, of course, range from $-\infty$ to $+\infty$), this is only likely to be a problem for small sample sizes; for the kind of sample size found in most randomized controlled trials, the differences between the confidence intervals for a risk difference based on the binomial distribution and on Cheung's modified OLS method will be numerically too small to be of consequence.

The problem is slightly more acute, however, for estimating individual group risk levels, as confidence intervals for risks close to the parameter boundary (i.e. close to 0% and

100%) can exceed the parameter boundary. Thus, negative lower confidence limits are possible for risk levels close to 0%, and upper confidence limits in excess of 100% are possible for risk levels close to 100%; vigilance is required to spot and appropriately adjust these should they occur.

In summary, therefore, the binomial model with an identity link function is the method of first choice for estimating risk / efficacy differences, provided the model converges. If, as happens worryingly frequently, the model fails to converge, this problem may be overcome by using the Copy method – but if convergence still fails when the number of copies has reached 10, the binomial model should be considered to have failed and the more reliable Cheung modified OLS method should be used.

Because of its considerably greater reliability, Cheung's modified OLS method was used exclusively throughout the next Chapter, which looks at methods for handling missing binary outcomes in the context of a randomized controlled trial.

Chapter 5 : An evaluation of methods for handling missing binary outcome values using imputation modeling

5.1 Mathematical approaches for imputing binary outcomes

Two different, but related, approaches were used in this Chapter to impute missing binary outcome values when fitting substantive models.

- In one set of analyses, missing outcomes were imputed as binary values using the Stata command: *mi impute logit.* With this command, the imputed outcome values are constrained to take the values 0 or 1 only.
- Then, in a second (otherwise identical) set of analyses, missing outcomes were imputed on a continuous scale using the Stata command: *mi impute regress*. With this command, the imputed outcome values are not constrained to be 0 or 1 but can take any value between these two boundary values.

Both sets of analyses were carried out using Cheung's modified OLS method to obtain adjusted efficacy differences; this approach allows efficacy (risk) differences to be computed whether the outcome is binary or continuous since the approach employs the OLS regression technique to estimate the model parameters and then adjusts the standard error estimates for so called "model errors***". In practice, this means that, in cases where the imputed values were on a continuous scale in the fitted regression model, the outcome variable contained both the observed binary outcomes and the imputed continuous outcome values.***

In the remainder of this Chapter, the findings obtained when the outcome was treated as binary are presented first; the results for data models that used continuous imputed outcomes then follow.

5.2 Results for missing data simulations with binary imputed outcomes

This section presents the findings of the statistical analyses of simulated data sets containing missing binary outcome values generated using three different missing data mechanism assumptions, namely: MCAR, MAR and MNAR. Under all three assumptions and for several different effect size scenarios, missing outcomes were imputed as *binary* variables.

Simulated data sets were generated for the following efficacy rate differences (effect sizes) under each of the three missing data mechanism assumptions:

- 60% efficacy in group A versus 85% efficacy in group B;
- 60% efficacy in group A versus 98% efficacy in group B;
- 95% efficacy in group A versus 98% efficacy in group B.

As detailed in the methodology section in Chapter 3, analyses are reported for 5%, 15% and 30% missing rates for each of the above efficacy scenarios.

5.2.1 Missing At Random (MAR) scenarios

5.2.1.1 Efficacy rates 85% vs. 60%

This scenario was purposively chosen such that both efficacy rates were away from the boundary values and to ensure that there was a substantial efficacy difference (effect size). The results of these analyses are presented in Table 5.1.

Predictably, as the proportion of missing data increased (and hence as the sample size effectively decreased), the effect size estimates from the complete case (CC) became increasingly inefficient (i.e. the standard error of this estimate became larger).

Table 5.1: Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.950	0.000
5% missing					
Complete Case	0.250 (0.063)	0.127	0.372	0.946	0.000
MI: wt, hb, age, para	0.238 (0.063)	0.116	0.361	0.955	-0.012
MI: hb, age, para	0.235 (0.063)	0.112	0.358	0.955	-0.015
MI: hb, age, para, **group**	0.251 (0.063)	0.128	0.374	0.946	+0.001
MI: wt, hb, age, para, **group**	0.250 (0.063)	0.127	0.373	0.948	0.000
15% missing					
Complete Case	0.250 (0.066)	0.120	0.380	0.945	0.000
MI: wt, hb, age, para	0.213 (0.066)	0.083	0.343	0.945	-0.037
MI: hb, age, para	0.211 (0.066)	0.081	0.341	0.944	-0.039
MI: hb, age, para, **group**	0.251 (0.066)	0.121	0.381	0.946	+0.001
MI: wt, hb, age, para, **group**	0.250 (0.066)	0.119	0.380	0.949	0.000
30% missing					
Complete Case	0.250 (0.073)	0.106	0.393	0.948	0.000
MI: wt, hb, age, para	0.174 (0.071)	0.034	0.313	0.884	-0.076
MI: hb, age, para	0.173 (0.071)	0.034	0.313	0.876	-0.077
MI: hb, age, para, **group**	0.248 (0.073)	0.104	0.391	0.941	-0.002
MI: wt, hb, age, para, **group**	0.246 (0.073)	0.103	0.390	0.946	-0.004

Much less predictably, however, exactly the same trends occurred for all of the imputation models evaluated. For small to moderate amounts of missing outcome data, this finding held irrespective of whether the model was correctly specified or misspecified – but, surprisingly, with 30% of outcomes missing, efficiency was slightly

better when the model was misspecified (i.e. when treatment group was not included in the model).

The estimates of adjusted efficacy difference were unbiased for all missing value levels when a complete case (CC) analysis was performed. Only small amounts of bias were detected for those imputation models which included group; for those models that did not include group (i.e. for misspecified imputation models), however, the estimates were markedly biased, and the degree of bias increased as the proportion of missing outcome values increased.

Coverage was generally high for all models at all missing value levels, remaining above 0.941 (94.1%). The only (and notable) exception occurred when the proportion of missing outcomes reached 30%; in this situation, coverage for the misspecified models fell to around 88%, which is unacceptably low.

In this MAR scenario, imputation models not containing both of the variables wt and group are technically misspecified as it is these two variables that determine missingness. As expected, therefore, the model containing both wt and group performed well for all missing outcome configurations, providing estimates that were only fractionally biased and with high coverage.

Less expectedly, models including group but excluding wt performed as well as the model with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included.

In summary, although both the group and wt variables were correlated with missingness, the inclusion of group in the imputation models greatly improved the performance of the multiple imputation procedures and provided unbiased estimates of effect size, whereas the inclusion of weight did not improve performance and produced biased estimates of effect size. These findings appear to indicate that, for the estimation of effect size, if missingness is related to group membership, excluding this variable from the imputation process is critical and will produce biased estimates; however, provided group is included in the imputation process, the absence of other covariates or factors linked to missingness has relatively little impact on bias levels.

5.2.1.2 Efficacy rates 98% vs. 60%

This scenario was purposively chosen such that one efficacy rate was close to a boundary value and to ensure that there was a substantial efficacy difference (38%). This is a common scenario in anti-malaria treatment efficacy trials where a new drug is highly efficacious while the standard drug has low efficacy, possibly due to the development of resistance. The results of these analyses are presented in Table 5.2 below.

Table 5.2: Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.479	0.946	0.000
5% missing					
Complete Case	0.380 (0.053)	0.276	0.483	0.942	0.000
MI: wt, hb, age, para	0.360 (0.053)	0.256	0.465	0.939	-0.020
MI: hb, age, para	0.361 (0.053)	0.257	0.465	0.944	-0.019
MI: hb, age, para, **group**	0.382 (0.053)	0.278	0.486	0.952	+0.002
MI: wt, hb, age, para, **group**	0.382 (0.053)	0.278	0.486	0.952	+0.002
15% missing					
Complete Case	0.380 (0.056)	0.270	0.490	0.941	0.000
MI: wt, hb, age, para	0.323 (0.058)	0.210	0.435	0.851	-0.057
MI: hb, age, para	0.321 (0.058)	0.208	0.434	0.856	-0.059
MI: hb, age, para, **group**	0.380 (0.056)	0.270	0.491	0.948	0.000
MI: wt, hb, age, para, **group**	0.380 (0.056)	0.270	0.491	0.947	0.000
30% missing					
Complete Case	0.380 (0.062)	0.259	0.501	0.939	0.000
MI: wt, hb, age, para	0.264 (0.063)	0.140	0.388	0.567	-0.116
MI: hb, age, para	0.264 (0.063)	0.140	0.388	0.563	-0.116
MI: hb, age, para, **group**	0.378 (0.063)	0.255	0.501	0.945	-0.002
MI: wt, hb, age, para, **group**	0.377 (0.063)	0.254	0.500	0.947	-0.003

As for the previous scenario above, when the proportion of missing data was increased (and hence the sample size effectively decreased), the standard errors of the effect size estimates from the complete case (CC) analyses also increased and the effect size estimates became increasingly inefficient. The exact same trends were observed for all of the imputation models evaluated – but (possibly because of the very large effect size

being simulated in this scenario), the efficiency of the models did not appear to be affected by misspecification.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed. Only small amounts of bias were detected for those imputation models which included group - but again, as in the previous scenario, for those models that did not include group the effect size estimates were markedly biased, and the degree of bias increased as the proportion of missing outcome values increased.

Coverage was generally high (0.939 (93.1%) or greater) for all CC analyses and for all imputation models which included group as a factor. For misspecified models not including group, however, coverage fell to unacceptably low levels: just over 85% with 15% missing data and just under 57% with 30% missing outcomes.

Fully specified imputation models containing both wt and group performed well for all missing outcome configurations, providing estimates that were only fractionally biased and with high coverage. In line with the previous scenario, models including group but excluding wt performed as well as the model with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included.

In summary, these findings appear to confirm that, for the estimation of a large effect size even if one of the effect sizes is close to a boundary value, if missingness is related to group membership, excluding this variable from the imputation process will produce biased estimates – but, provided group is included in the imputation process, the absence of other covariates or factors linked to missingness has little impact on bias levels.

5.2.1.3 Efficacy rates 98% vs. 95%

This scenario was purposively chosen such that the efficacy rates in both groups were close to boundary values. This is also another common scenario in malaria efficacy studies in which both drugs may be highly efficacious. The results of these analyses are reported in Table 5.3 below.

All complete case (CC) analyses converged without any problem - but, even though the usually reliable Cheung's modified OLS method was used, a small number of imputed analyses failed to converge, a problem that became more frequent as the proportion of missing outcome values increased. However, this is extremely unlikely to be a problem with Cheung's method; a more plausible explanation is that, on (the relatively rare) occasions when both efficacy rates are close to the same boundary, the imputation method replaces all missing outcome values with the same predicted outcome value and so both efficacy estimates go to the boundary. In this case specifically, it is possible that some imputation analyses resulted in all outcome values being 1, so the effect size and its standard error were both zero, causing even the OLS method to fail.

Table 5.3: Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets *	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data	5000	0.030 (0.026)	-0.021	0.079	0.950	0.000
5% missing						
Complete Case	5000	0.030 (0.026)	-0.022	0.081	0.940	0.000
MI: wt, hb, age, para	4981	0.029 (0.027)	-0.024	0.082	0.957	-0.001
MI: hb, age, para	4987	0.029 (0.027)	-0.024	0.081	0.953	-0.001
MI: hb, age, para, **group**	4988	0.031 (0.027)	-0.022	0.084	0.954	+0.001
MI: wt, hb, age, para, **group**	4990	0.031 (0.027)	-0.022	0.085	0.956	+0.001
15% missing						
Complete Case	5000	0.030 (0.028)	-0.024	0.084	0.942	0.000
MI: wt, hb, age, para	4970	0.026 (0.029)	-0.032	0.084	0.977	-0.004
MI: hb, age, para	4972	0.026 (0.029)	-0.031	0.083	0.973	-0.004
MI: hb, age, para, **group**	4975	0.031 (0.030)	-0.028	0.091	0.957	+0.001
MI: wt, hb, age, para, **group**	4977	0.032 (0.031)	-0.029	0.092	0.957	+0.002
30% missing						
Complete Case	5000	0.030 (0.030)	-0.029	0.090	0.937	0.000
MI: wt, hb, age, para	4887	0.022 (0.034)	-0.044	0.088	0.987	0.008
MI: hb, age, para	4933	0.021 (0.033)	-0.043	0.086	0.563	-0.009
MI: hb, age, para, **group**	4933	0.032 (0.037)	-0.041	0.105	0.945	+0.002
MI: wt, hb, age, para, **group**	4913	0.033 (0.038)	-0.042	0.107	0.947	+0.003

***: number of data sets for which convergent analysis was achieved.**

The sample size for these exploratory imputations was set at the same level as the original RCT. Clearly, this sample size was inadequate to detect an effect size as small as 0.030 (3%) as all of the analyses conducted under this scenario returned statistically non-significant findings (i.e. the 95% confidence intervals for all of the effect size estimates spanned zero). This absence of technical statistical significance does not affect the validity of either the analyses presented or the inferences drawn from them as the primary objective was to ascertain the influence of missing outcome values (and the methods used to handle these) on the *estimate* of effect size (difference in efficacy levels between the two study groups).

Again, increasing the proportion of missing data (and hence decreasing the effective sample size) increased the standard errors of the effect size estimates and decreased statistical efficiency in all analyses, although this trend was slightly less marked for the CC analyses than for the imputed analyses. The decrease in efficiency was slightly greater when group was included in the imputed model analyses than when this key missingness variable was excluded.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed, and only small levels of bias were detected for those imputation models which included group. As in both previous scenarios, for those models that did not include group the bias in the effect size estimates was markedly greater. When 30% of outcomes were missing, the imputation models produced bias

levels of 0.008 and 0.009; although numerically small, as the efficacy difference is only 0.030, this represents non-ignorable bias levels of 27% and 30% respectively.

For CC analyses, coverage decreased fractionally as the proportion of missing outcomes increased but, even with 30% missing, coverage was a respectable 0.937 (93.7%). Unexpectedly, however, coverage appeared to increase for the imputed model analyses as the proportion of missing outcomes increased. This was likely to be a consequence of the imputation process pushing both efficacy rate estimates close to the boundary for some models (and hence decreasing the estimate of efficacy difference for such models).

In summary, in this MAR situation in which missingness was related to group and wt, fully specified imputation models containing both of these variables performed well for all missing outcome configurations, providing estimates that were only fractionally biased and with high coverage; furthermore, imputed models including group but excluding wt performed as well as similar models with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included.

5.2.1.4 Imputing MAR binary outcomes with binary estimates - summary

In this situation, CC analyses performed as well, and often better, than imputed model analyses, consistently producing unbiased estimates of effect size. Some degree of efficiency was lost as the percentage of missing outcomes increased, due to the resulting

decrease in effective sample size, but this was no worse than found in the imputed model analyses.

No convergence problems were detected with the CC analyses. Convergence problems were experienced, however, when imputation models were used as an alternative method for handling missing outcomes (missing binary outcomes being replaced by binary imputation "estimates") in the situation where both efficacy (risk) levels were close to the parameter boundaries, due to all imputed values being allocated to the same outcome value resulting in zero standard errors for the effect size estimate.

If missingness in a binary outcome is MAR and related to study group membership, excluding this variable from the imputation process appears to produce biased estimates. However, if missingness is related also to other factors or covariates, the absence of these in the imputation model appears to have little impact on bias levels for the effect size estimate *provided group is included in the model* even if one of the effect sizes is close to a boundary value.

5.2.2 Missing Completely At Random (MCAR) scenarios

In this section exactly the same set of efficacy rates are presented as in the previous section but the assumption was made that "data was missing completely at random".

5.2.2.1 Efficacy rates 85% vs. 60%

The results of these analyses are presented in Table 5.4.

Table 5.4: Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.946	0.000
5% missing					
Complete Case	0.250 (0.062)	0.127	0.372	0.946	0.000
MI: wt, hb, age, para	0.238 (0.063)	0.115	0.361	0.955	-0.012
MI: hb, age, para	0.235 (0.063)	0.111	0.358	0.953	-0.015
MI: hb, age, para, **group**	0.250 (0.063)	0.128	0.373	0.948	0.000
MI: wt, hb, age, para, **group**	0.250 (0.063)	0.127	0.372	0.946	0.000
15% missing					
Complete Case	0.250 (0.066)	0.120	0.379	0.946	0.000
MI: wt, hb, age, para	0.212 (0.067)	0.082	0.343	0.941	-0.038
MI: hb, age, para	0.211 (0.067)	0.081	0.342	0.949	-0.037
MI: hb, age, para, **group**	0.250 (0.066)	0.120	0.380	0.941	0.000
MI: wt, hb, age, para, **group**	0.250 (0.066)	0.119	0.379	0.947	0.000
30% missing					
Complete Case	0.250 (0.073)	0.107	0.393	0.946	0.000
MI: wt, hb, age, para	0.174 (0.072)	0.033	0.314	0.888	-0.076
MI: hb, age, para	0.174 (0.072)	0.034	0.315	0.880	-0.076
MI: hb, age, para, **group**	0.248 (0.073)	0.104	0.391	0.951	-0.002
MI: wt, hb, age, para, **group**	0.247 (0.073)	0.103	0.390	0.945	-0.003

Exactly as observed in the MAR case for this efficacy difference scenario, as the proportion of missing data increased (and so as the effective sample size decreased), the

effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger).

Like in the MAR case, when outcomes were missing MCAR, no bias was found in the effect size estimates both for the CC analyses *and* for those imputed model analyses that included study group membership in the imputation process. This is expected from theory for the CC analyses since the remaining subjects constitute a random sample of the target population, but is perhaps less predictable for the imputed models involving group membership.

Imputation models that failed to include group in the imputation process produced (negative) bias levels of around 15% when 15% of outcomes were missing and of around 30% when the proportion of missing outcomes was as high as 30%.

The MAR and MCAR analyses were very similar with respect to coverage levels. In the MCAR situation, coverage was generally high for all models at all missing value levels, again remaining above 0.941 (94.1%), but with the notable exception of when the proportion of missing outcomes reached 30%; in this situation, coverage for imputation models with group excluded fell to around 88%, which is identical to that observed in the MAR situation and again unacceptably low.

In summary, even when missingness in a (binary) outcome measure is MCAR, the inclusion of study group membership in the imputation models appears to greatly improve the performance of the multiple imputation procedure and provides unbiased estimates of effect size even with 30% of outcomes unrecorded. Excluding group from the imputation process, however, appears to produce marked levels of bias (tending to under-estimate the true effect size).

5.2.2.2 Efficacy rates 98% vs. 60%

The results of these analyses are presented in Table 5.5 below.

Table 5.5: Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.480	0.950	0.000
5% missing					
Complete Case	0.379 (0.052)	0.276	0.482	0.942	-0.001
wt, hb, age, para	0.360 (0.054)	0.255	0.465	0.941	-0.020
MI: hb, age, para	0.361 (0.054)	0.256	0.466	0.945	-0.019
MI: hb, age, para, **group**	0.379 (0.053)	0.276	0.482	0.945	-0.001
MI: wt, hb, age, para, **group**	0.382 (0.053)	0.278	0.485	0.950	+0.002
15% missing					
Complete Case	0.380 (0.056)	0.271	0.489	0.946	0.000
MI: wt, hb, age, para	0.323 (0.058)	0.209	0.436	0.870	-0.057
MI: hb, age, para	0.321 (0.058)	0.207	0.435	0.861	-0.059
MI: hb, age, para, **group**	0.380 (0.056)	0.270	0.489	0.941	0.000
MI: wt, hb, age, para, **group**	0.380 (0.056)	0.270	0.489	0.948	0.000
30% missing					
Complete Case	0.380 (0.061)	0.260	0.500	0.944	0.000
MI: wt, hb, age, para	0.264 (0.064)	0.139	0.389	0.577	-0.116
MI: hb, age, para	0.264 (0.064)	0.139	0.389	0.574	-0.116
MI: hb, age, para, **group**	0.378 (0.062)	0.256	0.500	0.948	-0.002
MI: wt, hb, age, para, **group**	0.377 (0.062)	0.255	0.498	0.944	-0.003

The pattern of results is very similar in this as in the previous scenario:

- As the proportion of missing data increased (and effective sample size decreased), the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger) at identical rates.

- Little or no bias was found in the effect size estimates both for the CC analyses *and* for those imputed model analyses that included study group membership in the imputation process.
- Those imputation models that did not include group in the imputation process again produced (negative) bias levels of around 15% when 15% of outcomes were missing and of around 30% when the proportion of missing outcomes was as high as 30%.
- Coverage was generally high for all models at all missing value levels, remaining above 0.941 (94.1%) while the proportion of missing outcomes was no greater than 5%. However, when the proportion of missing outcomes reached 15%, coverage for the imputation models excluding group fell to just under 87%, and when the proportion of missing outcomes reached 30%, coverage for these same models fell even further to just under 58%.

In summary, as before, when missingness in a (binary) outcome measure is MCAR, the inclusion of study group membership in the imputation models appears to greatly improve the performance of the multiple imputation procedure and provides unbiased estimates of effect size even with 30% of outcomes unrecorded. Excluding group from the imputation process, however, appears to produce marked levels of bias (tending to under-estimate the true effect size) which increase as the proportion of missing outcome values increases.

5.2.2.3 Efficacy rates 98% vs. 95% efficacy

The results of these analyses are presented in Table 5.6.

Table 5.6: Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets *	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data		0.030 (0.026)	-0.020	0.080	0.939	0.000
5% missing						
Complete Case	5000	0.030 (0.026)	-0.021	0.081	0.940	0.000
MI: wt, hb, age, para	4980	0.029 (0.027)	-0.024	0.082	0.955	-0.001
MI: hb, age, para	4983	0.028 (0.027)	-0.024	0.081	0.957	-0.002
MI: hb, age, para, **group**	4997	0.030 (0.027)	-0.023	0.083	0.949	0.000
MI: wt, hb, age, para, **group**	4986	0.031 (0.027)	-0.023	0.084	0.948	+0.001
15% missing						
Complete Case	5000	0.030 (0.028)	-0.024	0.084	0.944	0.000
MI: wt, hb, age, para	4964	0.026 (0.029)	-0.032	0.083	0.973	-0.004
MI: hb, age, para	4984	0.025 (0.029)	-0.032	0.082	0.972	-0.005
MI: hb, age, para, **group**	4986	0.031 (0.030)	-0.029	0.090	0.956	+0.001
MI: wt, hb, age, para, **group**	4974	0.030 (0.031)	-0.030	0.091	0.960	0.000
30% missing						
Complete Case	5000	0.030 (0.030)	-0.029	0.089	0.940	0.000
MI: wt, hb, age, para	4909	0.022 (0.034)	-0.044	0.088	0.985	-0.008
MI: hb, age, para	4933	0.021 (0.033)	-0.044	0.085	0.981	-0.009
MI: hb, age, para, **group**	4943	0.032 (0.037)	-0.041	0.104	0.970	+0.002

MI: wt, hb, age, para, **group**	4933	0.031 (0.038)	-0.043 0.106	0.975	+0.001

***: number of data sets for which convergent analysis was achieved.**

As in the MAR situation reported earlier, all complete case (CC) analyses converged without any problem but again a small number of imputed analyses failed to converge and this problem increased as the proportion of missing outcome values increased.

The sample size again proved inadequate to detect an effect size as small as 0.030 (3%). All of the analyses returned statistically non-significant findings.

As the proportion of missing data increased (and effective sample size decreased), the standard errors of the effect size estimates increased and the efficiency of all analyses decreased; this trend was again less marked for the CC analyses than for the imputed analyses. The decrease in efficiency was slightly greater when group was included in the imputed model analyses.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed, and only small levels of bias were detected for those imputation models which included group. As in both previous scenarios, the bias in the effect size estimates was markedly greater using imputation models that did not include group, reaching ~15% when 15% of outcomes were missing and ~30% when 30% of outcomes were missing.

In this situation, coverage was worst (0.939 compared to the set nominal level of 0.950) when the full data set was analysed. When missing outcomes were introduced, coverage ranged from 0.940 to 0.985; these higher values were again attributed to the imputation process pushing both efficacy rate estimates close to the boundary for some models.

In summary, in this scenario CC analyses provided unbiased effect size estimates with good coverage and efficiency. Imputation models containing group membership in the imputation process returned estimates that were only fractionally biased and with high coverage, although efficiency was observed to fall as the proportion of missing outcomes increased. Imputed models that did not include group in the imputation process were moderately efficient but produced biased estimates of effect size (the level of bias increasing with the proportion of missing outcomes).

5.2.2.4 Imputing MCAR binary outcomes with binary estimates - summary

As in the MAR situation, CC analyses performed as well, and often better, than imputed model analyses, consistently producing unbiased estimates of effect size. Some degree of efficiency was inevitably lost as the percentage of missing outcomes increased, due to the resulting decrease in effective sample size, but this was no worse than found in the imputed model analyses.

No convergence problems were detected with the CC analyses, but these were experienced when imputation models were used as an alternative method for handling

missing outcomes (missing binary outcomes being replaced by binary imputation "estimates") in the situation where both efficacy (risk) levels were close to the parameter boundaries, due to all imputed values being allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

In summary, even when missing binary outcomes are MCAR (i.e. when missingness is effectively wholly random), excluding study group membership from the imputation process appears to produce biased estimates.

5.2.3 Missing Not At Random (MNAR) scenarios

In this section the same set of efficacy rates are presented as in the previous sections, but the assumption is now made that "data is missing not at random".

Specifically, missingness was simulated to be related to outcome. Subjects for whom the treatment was successful (i.e. was coded 1) were more likely to have a missing outcome than subjects for whom the treatment was deemed to have failed (i.e. was coded 0). This resulted in differential missingness as the group with the higher efficacy rate was likely to lose more people with success outcomes than the inferior efficacy group.

The CC analysis, an imputation model containing all covariates (age, hb, wt and para) plus group, and an imputation model containing age, hb, wt and para but without group were all assessed. The objective was to assess whether complete case analysis and

imputation models that included group continued to perform better than imputation models that excluded group, knowing quite well that none of these models was correct for these missing data scenarios.

5.2.3.1 Efficacy rates 85% vs. 60%

The results of these analyses are presented in Table 5.7.

Table 5.7: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.950	0.000
5% missing					
Complete Case	0.258 (0.063)	0.134	0.381	0.942	+0.008
MI: wt, hb, age, para	0.245 (0.064)	0.120	0.370	0.957	-0.005
MI: wt, hb, age, para, **group**	0.258 (0.063)	0.133	0.382	0.947	+0.008
15% missing					
Complete Case	0.274 (0.068)	0.140	0.408	0.932	+0.024
MI: wt, hb, age, para	0.233 (0.070)	0.097	0.370	0. 972	-0.017
MI: wt, hb, age, para, **group**	0.273 (0.069)	0.138	0.407	0.933	+0.023
30% missing					
Complete Case	0.298 (0.078)	0.145	0.452	0.895	+0.048
MI: wt, hb, age, para	0.207 (0.078)	0.054	0.359	0.975	-0.043
MI: wt, hb, age, para, **group**	0.294 (0.078)	0.141	0.447	0.897	+0.044

As observed in both the MAR and MCAR cases for this efficacy difference scenario, when the proportion of missing data was increased (and hence the effective sample size reduced), the effect size estimates from both the CC and imputed model analyses became increasingly, and unacceptably, inefficient (i.e. the standard error of this estimate became larger).

Unlike in both the MAR and the MCAR cases, when outcomes were MNAR there was some degree of bias in the effect size estimates from the CC analyses *and* from both imputed model analyses (i.e. irrespective of whether group was used in the imputation process). Even more surprisingly, for the multiple imputation analyses:

- models that included group in the imputation process tended to be more biased than models that excluded group (although the differences were numerically quite small);
- models that included group exhibited progressively more *positive* bias as the proportion of missing outcome values increased while models that excluded group exhibited progressively more *negative* bias.

In summary, both the CC and multiple imputation models produced biased estimates of effect size. The exclusion of study group membership in the imputation models maintained coverage levels but produced negatively biased estimates of effect size. Including study group membership produced virtually identical results to the CC analysis, with coverage markedly reduced and effect size consistently over-estimated.

5.2.3.2 Efficacy rates 98% vs. 60%

The results of these analyses are presented in Table 5.8.

Table 5.8: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.479	0.950	0.000
5% missing					
Complete Case	0.394 (0.053)	0.291	0.497	0.938	+0.014
MI: wt, hb, age, para	0.374 (0.054)	0.267	0.480	0.956	-0.006
MI: wt, hb, age, para, **group**	0.396 (0.053)	0.292	0.499	0.946	+0.016
15% missing					
Complete Case	0.426 (0.056)	0.317	0.536	0.863	+0.046
MI: wt, hb, age, para	0.361 (0.061)	0.242	0.481	0.973	-0.019
MI: wt, hb, age, para, **group**	0.426 (0.056)	0.315	0.536	0.870	+0.046
30% missing					
Complete Case	0.484 (0.062)	0.364	0.605	0.597	+0.104
MI: wt, hb, age, para	0.333 (0.071)	0.194	0.472	0.973	-0.047
MI: wt, hb, age, para, **group**	0.478 (0.063)	0.356	0.601	0.641	+0.098

As for all the previous scenarios, as the proportion of missing data increased, the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger). This

trend was slightly more pronounced for the multiple imputation models with group membership excluded from the imputation process.

As in the previous comparison (85% vs. 60% efficacy), the CC analyses and both imputation models (i.e. irrespective of whether group was included in the imputation process) produced biased effect size estimates. Independent of the proportion of missing outcome values, compared with the imputed models that excluded group, the CC analyses and the imputed models that included group:

- were numerically more biased
- were positively rather than negatively biased
- were less efficient - coverage rates were poor (86.3% and 87.0% respectively) at 15% missing outcomes, and wholly unacceptable at the 30% missing rate (59.7% and 64.1% respectively).

In summary, both the CC and multiple imputation models produced biased estimates of effect size. Again, the exclusion of study group membership in the imputation models maintained coverage levels but produced negatively biased estimates of effect size, while including study group membership in the multiple imputation process produced virtually identical results to the CC analysis, with coverage markedly reduced and effect size consistently over-estimated.

5.2.3.3 Efficacy rates 98% vs. 95%

The results of these analyses are presented in Table 5.9.

Table 5.9: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets*	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data	5000	0.030 (0.026)	-0.021	0.079	0.935	0.000
5% missing						
Complete Case	5000	0.031 (0.027)	-0.021	0.084	0.947	+0.001
MI: wt, hb, age, para	4984	0.030 (0.027)	-0.024	0.084	0.964	0.000
MI: wt, hb, age, para, **group**	4991	0.031 (0.028)	-0.023	0.086	0.954	+0.001
15% missing						
Complete Case	5000	0.035 (0.030)	-0.024	0.094	0.951	+0.005
MI: wt, hb, age, para	4989	0.029 (0.032)	-0.032	0.091	0.981	-0.001
MI: wt, hb, age, para, **group**	4987	0.036 (0.032)	-0.028	0.099	0.962	+0.006
30% missing						
Complete Case	5000	0.042 (0.036)	-0.029	0.113	0.950	+0.012
MI: wt, hb, age, para	4985	0.029 (0.038)	-0.045	0.103	0.994	-0.001
MI: wt, hb, age, para, **group**	4990	0.043 (0.041)	-0.039	0.124	0.961	+0.013

***: number of data sets for which convergent analysis was achieved.**

As in the previous situations reported in which both efficacy rates were close to a boundary value, all complete case (CC) analyses converged without any problem - but

again a small number of imputed analyses failed to converge and this problem increased as the proportion of missing outcome values increased.

Like the MAR and MCAR cases, the sample size proved insufficient to detect an effect size as small as 0.030 (3%). All of the analyses again returned statistically non-significant findings.

Again, increasing the proportion of missing data increased the standard errors of the effect size estimates and decreased statistical efficiency in all analyses. The decrease in efficiency was slightly greater in the imputed model analyses than the complete case analyses especially for the moderate to large missing rates.

It is interesting, and perhaps surprising, to note that the imputed model analyses that excluded group produced estimates of effect size that were only very slightly biased. The CC analyses and imputed model analyses that included group produced almost identically biased estimates of effect size, with bias increasing with the proportion of missing outcomes. When 30% of outcomes were missing, the CC analyses and the imputation models that included group produced bias levels of 0.012 and 0.013 respectively; although numerically small, as the efficacy difference is only 0.030, this represents non-ignorable bias levels of 40% and 43% respectively.

Coverage was at least 94.7% for all scenarios and all levels of missing data.

In summary, when outcomes are MNAR and both groups have efficacy levels close to the parameter boundary, the CC analyses and multiple imputation analyses that included group in the imputation process produced very similar positively biased estimates of effect size, the level of bias being very high when 30% of outcomes were missing. Multiple imputation models that excluded group in the imputation process produced estimates of effect size that were only very marginally biased.

5.2.3.4 Imputing MNAR binary outcomes with binary estimates - summary

Unlike in the MAR and MCAR situations, both CC analyses and the imputed model analyses (irrespective of whether group was included or not) produced biased estimates of effect size.

Consistent with the findings from all the other scenarios considered above, imputation models in which group was excluded from the imputation process produced estimates of effect size that were negatively biased (i.e. that were biased toward the null hypothesis). However, the really striking finding was that the estimates from the CC analyses and from imputation models that included group in the imputation process produced estimates of effect size that were positively biased (i.e. that were consistently biased away from the null hypothesis) – and that the degree of bias was similar for both models.

Inevitably, some degree of efficiency was lost as the percentage of missing outcomes increased in both the CC analyses and the imputed model analyses irrespective of whether group was included or not. Coverage tended to be higher than the expected 95% using multiple imputation models excluding group.

No convergence problems were detected with the CC analyses, but these were experienced when imputation models were used as an alternative method for handling missing outcomes (missing binary outcomes being replaced by binary imputation "estimates") in the situation where both efficacy (risk) levels were close to the parameter boundaries, due to all imputed values being allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

5.3 Results for missing data simulations with continuous imputed outcomes

This section presents the findings of the statistical analyses of simulated data sets containing missing binary outcome values generated using three different missing data mechanism assumptions, namely: MCAR, MAR and MNAR. Under all three assumptions and for several different effect size scenarios, missing outcomes were now imputed as *continuous* variables. In the analyses of substantive models the outcome variable contained both the observed binary outcomes and the continuous imputed outcomes.

As in section 5.2, simulated data sets were generated for the following efficacy rate differences (effect sizes) under each of the three missing data mechanism assumptions:

- 60% efficacy in group A versus 85% efficacy in group B;
- 60% efficacy in group A versus 98% efficacy in group B;
- 95% efficacy in group A versus 98% efficacy in group B.

As detailed in the methodology section in Chapter 3, analyses are reported for 5%, 15% and 30% missing rates for each of the above efficacy scenarios. The main aim of these scenarios was to establish whether imputing missing outcomes as a continuous variable has different statistical implications compared to the intuitively more conventional process of imputing missing outcomes as a binary variable.

5.3.1 Missing At Random (MAR) scenarios

5.3.1.1 Efficacy rates 85% vs. 60%

The results of these analyses are presented in Table 5.10.

Table 5.10: Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.946	0.000
5% missing					
Complete Case	0.250 (0.063)	0.127	0.373	0.952	0.000
MI: wt, hb, age, para	0.238 (0.063)	0.114	0.361	0.955	-0.012
MI: hb, age, para	0.238 (0.063)	0.114	0.362	0.947	-0.012
MI: hb, age, para, **group**	0.251 (0.063)	0.128	0.374	0.949	+0.001
MI: wt, hb, age, para, **group**	0.251 (0.063)	0.128	0.373	0.951	+0.001
15% missing					
Complete Case	0.250 (0.066)	0.120	0.380	0.945	0.000
MI: wt, hb, age, para	0.212 (0.068)	0.080	0.345	0.946	-0.038
MI: hb, age, para	0.212 (0.067)	0.080	0.344	0.947	-0.038
MI: hb, age, para, **group**	0.251 (0.067)	0.119	0.383	0.946	+0.001
MI: wt, hb, age, para, **group**	0.250 (0.067)	0.119	0.382	0.945	0.000
30% missing					
Complete Case	0.250 (0.073)	0.105	0.393	0.950	0.000
MI: wt, hb, age, para	0.174 (0.073)	0.031	0.318	0.883	-0.076
MI: hb, age, para	0.173 (0.073)	0.030	0.317	0.882	-0.077
MI: hb, age, para, **group**	0.251 (0.075)	0.104	0.398	0.939	+0.001
MI: wt, hb, age, para, **group**	0.249 (0.075)	0.100	0.396	0.938	-0.001

These findings are virtually identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.1.1).

As the proportion of missing data increased, the effect size estimates from the complete case (CC) became increasingly inefficient (i.e. the standard error of this estimate became larger); exactly the same trends occurred for all of the imputation models evaluated. For small (5%) to moderate (15%) amounts of missing outcome data, these findings held irrespective of whether the model was correctly specified or misspecified.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when a complete case (CC) analysis was performed. As for imputed binary outcomes, only small amounts of bias were detected for those imputation models which included group; this small degree of bias was positive when 5% and 15% of outcomes were missing, but was negative when 30% of outcomes were not available. For those imputation models that did not include group (i.e. for misspecified imputation models), however, the estimates were markedly biased, and the degree of bias increased as the proportion of missing outcome values increased.

Coverage was generally high for all models at all missing value levels, remaining above 0.945 (94.5%) for small to moderate missing rates. Consistent with earlier findings for the binary imputed outcomes, when the proportion of missing outcomes reached 30%, coverage for the misspecified models fell to around 88%, which is unacceptably low.

As expected, the model containing both wt and group performed well for all missing outcome configurations, providing estimates that were only fractionally biased and with generally acceptable high coverage.

Less expectedly, models including group but excluding wt performed as well as the model with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included, agreeing with the earlier findings on the binary imputed outcomes.

In summary, therefore, although both the group and wt variables were correlated with missingness, the inclusion of group in the imputation models greatly improved the performance of the multiple imputation procedures and provided unbiased estimates of effect size, whereas the inclusion of wt did not improve performance and produced biased estimates of effect size. These findings appear to indicate that, for the estimation of effect size, if missingness is related to group membership, excluding this variable from the imputation process is critical and will produce biased estimates; however, provided group is included in the imputation process, the absence of other covariates or factors associated with missingness has relatively little impact on bias levels. These findings are consistent with those considered earlier where the missing outcomes were imputed as a binary variable.

5.3.1.2 Efficacy rates 98% vs. 60%

The results of these analyses are presented in Table 5.11. These findings are nearly identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.1.2).

As in the previous scenario considered above, when the proportion of missing data was increased, the standard errors of the effect size estimates from the complete case (CC) analyses also increased and the effect size estimates became increasingly inefficient. The exact same trends were observed for binary imputed outcomes of the imputation models evaluated – the efficiency of the models appear to be affected by misspecification with standard errors being higher than both the CC and the imputed models with group in them.

Table 5.11 Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.480	0.950	0.000
5% missing					
Complete Case	0.380 (0.053)	0.276	0.484	0.952	0.000
MI: wt, hb, age, para	0.361 (0.054)	0.256	0.466	0.945	-0.019
MI: hb, age, para	0.361 (0.054)	0.256	0.466	0.938	-0.019
MI: hb, age, para, **group**	0.380 (0.052)	0.278	0.483	0.946	0.000
MI: wt, hb, age, para, **group**	0.380 (0.052)	0.277	0.483	0.943	0.000
15% missing					
Complete Case	0.380 (0.056)	0.270	0.490	0.941	0.000
MI: wt, hb, age, para	0.323 (0.059)	0.208	0.438	0.857	-0.057
MI: hb, age, para	0.322 (0.058)	0.207	0.436	0.859	-0.058
MI: hb, age, para, **group**	0.381 (0.056)	0.271	0.491	0.935	+0.001
MI: wt, hb, age, para, **group**	0.380 (0.056)	0.270	0.490	0.939	0.000
30% missing					
Complete Case	0.380 (0.062)	0.259	0.501	0.952	0.000
MI: wt, hb, age, para	0.265 (0.065)	0.137	0.392	0.590	-0.115
MI: hb, age, para	0.263 (0.065)	0.136	0.390	0.573	-0.117
MI: hb, age, para, **group**	0.381 (0.063)	0.258	0.503	0.937	+0.001
MI: wt, hb, age, para, **group**	0.378 (0.063)	0.255	0.502	0.932	-0.002

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed. The estimates of adjusted efficacy difference were also unbiased for all missing value levels when MI imputed models with both group and wt were performed. Only small amounts of bias were detected for those imputation

models which included group. This small bias could go in either direction where it occurred - but again, for those models that did not include group the effect size estimates were markedly biased, and the degree of bias increased as the proportion of missing outcome values increased.

Coverage was generally high (0.93.5 (93.5%) or greater) for all CC analyses and for all imputation models which included group as a factor. For misspecified models not including group, however, coverage fell to unacceptably low levels for 30% missing data to just over 86%.

Fully specified imputation models containing both wt and group performed well for all missing outcome configurations, providing estimates that were only unbiased and with high coverage. In line with the previous scenarios, models including group but excluding wt performed as well as the model with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included.

These findings appear to confirm that, for the estimation of a large effect size even if one of the effect sizes is close to a boundary value, if missingness is related to group membership, excluding this variable from the imputation process will produce biased estimates – but, provided group is included in the imputation process, the absence of

other covariates or factors linked to missingness has little impact on bias levels. This is again consistent with the findings for the binary imputed outcomes.

5.3.1.3 Efficacy rates 98% vs. 95%

The results of these analyses are reported in Table 5.12. These findings are almost identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.1.3).

Table 5.12: Estimated efficacy differences, coverage and bias for different proportions of missing MAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets *	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data	5000	0.030 (0.026)	-0.020	0.080	0.939	0.000
5% missing						
Complete Case	5000	0.030 (0.026)	-0.022	0.081	0.940	0.000
MI: wt, hb, age, para	4997	0.028 (0.026)	-0.024	0.080	0.947	-0.002
MI: hb, age, para	4997	0.028 (0.026)	-0.023	0.080	0.950	-0.002
MI: hb, age, para, **group**	4997	0.029 (0.026)	-0.022	0.081	0.940	-0.001
MI: wt, hb, age, para, **group**	4995	0.031 (0.026)	-0.021	0.082	0.937	+0.001
15% missing						
Complete Case	5000	0.030 (0.028)	-0.024	0.085	0.939	0.000
MI: wt, hb, age, para	4992	0.025 (0.028)	-0.029	0.080	0.950	-0.005
MI: hb, age, para	4987	0.025 (0.028)	-0.029	0.080	0.951	-0.005
MI: hb, age, para, **group**	4992	0.030 (0.028)	-0.025	0.085	0.933	0.000
MI: wt, hb, age, para, **group**	4989	0.031 (0.028)	-0.024	0.086	0.939	+0.001
30% missing						
Complete Case	5000	0.030 (0.030)	-0.030	0.090	0.940	0.000
MI: wt, hb, age, para	4969	0.021 (0.030)	-0.038	0.080	0.960	-0.009
MI: hb, age, para	4964	0.021 (0.030)	-0.038	0.080	0.961	-0.009
MI: hb, age, para, **group**	4969	0.030 (0.031)	-0.032	0.092	0.934	0.000
MI: wt, hb, age, para, **group**	4953	0.030 (0.032)	-0.032	0.093	0.928	0.000

***: number of data sets for which convergent analysis was achieved.**

All complete case (CC) analyses converged without any problem – but interestingly, as in the previous scenarios with both efficacy rates close to a boundary value, a small

number of imputed analyses failed to converge, and this problem increased as the proportion of missing outcome values increased. Again the most plausible explanation is that, on (the relatively rare) occasions when both efficacy rates are close to the same boundary, the imputation method replaces all missing outcome values with the same predicted outcome value and so both efficacy estimates go to the boundary. For an efficacy rate of 98% it is possible to have all imputed values estimated as "1" even if the imputation is on a continuous scale. In this situation, there is no variability in the outcome variable resulting in standard errors that are zero causing the estimation procedure to fail.

Closer examination of the results, however, indicated that the proportion of non-convergent models is smaller in this scenario than for the binary imputed outcomes considered in section 5.2.1.3. Clearly, and mathematically predictably, if missing outcomes are imputed on a continuous scale, even when the efficacy is very high, the probability of imputing all of the missing outcomes as being exactly 1 is reduced. So, imputing missing binary outcome using a continuous scale reduces the risk of the model failing to converge but affects the effect size estimates minimally.

As in the previous scenarios with both efficacy rates close to a boundary value, the sample size used in the simulation was insufficient to detect an effect size as small as 0.030 (3%) as all of the analyses conducted under this scenario returned statistically non-significant findings (i.e. the 95% confidence intervals for all of the effect size estimates spanned zero). Again, this absence of technical statistical significance does

not affect the validity of either the analyses presented or the inferences drawn from them as the primary objective was to ascertain the influence of missing outcome values (and the methods used to handle these) on the *estimate* of effect size (difference in efficacy levels between the two study groups).

Once again, increasing the proportion of missing data (and hence decreasing the effective sample size) increased the standard errors of the effect size estimates and decreased statistical efficiency in all analyses. This trend was slightly more marked for the analyses of the moderate to high missing outcome rates than for the small missing outcome rate analyses.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed and for those imputation models which included group and wt. Only small levels of bias were detected for those imputation models which included group but not wt. As in previous scenarios, for those models that did not include group the bias in the effect size estimates was markedly greater. When 30% of outcomes were missing, the imputation models produced bias levels of 0.009; although numerically small, as the efficacy difference is only 0.030, this represents non-ignorable bias levels 30%.

For CC analyses, coverage decreased fractionally as the proportion of missing outcomes increased but, even with 30% of outcomes missing, coverage was a respectable 0.94

(94%). Unexpectedly, however, coverage appeared to increase for the imputed model analyses that did not include group in the imputation process as the proportion of missing outcomes increased. As explained in the related scenarios above, this was likely to be a consequence of the imputation process pushing both efficacy rate estimates close to the boundary for some models (and hence decreasing the estimate of efficacy difference for such models).

Surprisingly, this was not the case with those imputed models that included group. For those imputations, coverage decreased fractionally as the proportion of missing outcomes increased but, even with 30% missing, coverage was 0.928 (92.8%).

Again in this MAR situation in which missingness was related to group and wt, fully specified imputation models containing both of these variables performed well for all missing outcome configurations, providing estimates that were only fractionally biased and with high coverage; furthermore, imputed models including group but excluding wt performed as well as similar models with both group and wt included, whereas models including wt but excluding group performed as badly as models with neither wt nor group included.

5.3.1.4 Imputing MAR binary outcomes with continuous estimates - summary

In this situation, exactly as observed when MAR binary outcomes were imputed using a binary scale, CC analyses performed as well, and often better, than imputed model

analyses, consistently producing unbiased estimates of effect size. Some degree of efficiency was lost as the percentage of missing outcomes was increased, due to the resulting decrease in effective sample size, but this was no worse than found in the imputed model analyses.

No convergence problems were detected with the CC analyses. Convergence problems were experienced, however, when imputation models were used as an alternative method for handling missing outcomes (missing binary outcomes being replaced by continuous imputation "estimates") in the situation where both efficacy (risk) levels were close to the parameter boundaries, probably due to all imputed values being allocated to the same outcome. However, imputing missing binary outcomes on a continuous scale appears to reduce the risk of all imputed values being set at the same value and hence the risk of the model not converging.

If missingness in a binary outcome is MAR and related to study group membership, excluding this variable from the imputation process appears to produce biased estimates. However, if missingness is related also to other factors or covariates, the absence of these in the imputation model appears to have little impact on bias levels for the effect size estimate *provided group is included in the model* even if one of the effect sizes is close to a boundary value. These findings are consistent with the earlier findings when the outcomes were imputed as binary

5.3.2 Missing Completely At Random (MCAR) scenarios

5.3.2.1 Efficacy rates 85% vs. 60%

The results of these analyses are presented in Table 5.13. These findings are generally identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.2.1).

Table 5.13 Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.946	0.000
5% missing					
Complete Case	0.250 (0.062)	0.127	0.372	0.948	0.000
MI: wt, hb, age, para	0.238 (0.063)	0.115	0.361	0.955	-0.012
MI: hb, age, para	0.237 (0.063)	0.113	0.361	0.958	-0.013
MI: hb, age, para, **group**	0.251 (0.063)	0.128	0.373	0.953	+0.001
MI: wt, hb, age, para, **group**	0.250 (0.063)	0.128	0.374	0.951	0.000
15% missing					
Complete Case	0.250 (0.066)	0.121	0.380	0.951	0.000
MI: wt, hb, age, para	0.212 (0.067)	0.080	0.344	0.944	-0.038
MI: hb, age, para	0.212 (0.067)	0.081	0.343	0.947	-0.038
MI: hb, age, para, **group**	0.250 (0.067)	0.119	0.381	0.941	0.000
MI: wt, hb, age, para, **group**	0.250 (0.067)	0.119	0.381	0.948	0.000
30% missing					
Complete Case	0.250 (0.073)	0.108	0.394	0.954	0.000
MI: wt, hb, age, para	0.173 (0.073)	0.030	0.316	0.890	-0.077
MI: hb, age, para	0.173 (0.073)	0.031	0.316	0.895	-0.077
MI: hb, age, para, **group**	0.249 (0.075)	0.103	0.395	0.945	-0.001
MI: wt, hb, age, para, **group**	0.250 (0.075)	0.103	0.396	0.950	0.000

Exactly as observed in the MAR case for this efficacy difference scenario, as the proportion of missing data increased, the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger).

Unlike in the MAR case, when outcomes were missing MCAR, no bias was found in the effect size estimates both for the CC analyses *and* for those imputed model analyses that included study group membership in the imputation process. Again this is expected from theory for the CC analyses since the remaining subjects are just a random sample of the target population, but is perhaps less predictable for the imputed models involving group membership.

Imputation models that did not include group in the imputation process produced (negative) bias levels of around 15% when 15% of outcomes were missing and of around 31% when the proportion of missing outcomes was as high as 30%.

The MAR and MCAR analyses were very similar with respect to coverage levels. In the MCAR situation, coverage was generally high for all models at all missing value levels, again remaining above 0.941 (94.1%), but with the notable exception of when the proportion of missing outcomes reached 30%; in this situation, coverage for imputation models with group excluded fell to around 89%, which is slightly higher than that observed in the MAR situation and again unacceptably low.

Thus, *in summary*, even when missingness in a (binary) outcome measure with continuous imputed is MCAR, the inclusion of study group membership in the imputation models appears to greatly improve the performance of the multiple imputation procedure and provides unbiased estimates of effect size even with 30% of

outcomes unrecorded. Excluding group from the imputation process, however, appears to produce marked levels of bias (tending to under-estimate the true effect size). The findings are similar to those whose outcomes were imputed as binary.

5.3.2.2 Efficacy rates 98% vs. 60%

The results of these analyses are presented in Table 5.14. These findings are virtually identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.2.2); furthermore, the pattern of results was very similar to the previous scenario (85% vs 60% efficacy).

Table 5.14: Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.480	0.950	0.000
5% missing					
Complete Case	0.380 (0.053)	0.277	0.483	0.950	0.000
MI: wt, hb, age, para	0.361 (0.054)	0.255	0.466	0.938	-0.019
MI: hb, age, para	0.360 (0.054)	0.255	0.466	0.934	-0.020
MI: hb, age, para, **group**	0.380 (0.053)	0.277	0.483	0.945	0.000
MI: wt, hb, age, para, **group**	0.380 (0.053)	0.277	0.484	0.950	0.000
15% missing					
Complete Case	0.380 (0.056)	0.271	0.489	0.946	0.000
MI: wt, hb, age, para	0.323 (0.059)	0.208	0.437	0.874	-0.057
MI: hb, age, para	0.322 (0.059)	0.208	0.437	0.868	-0.058
MI: hb, age, para, **group**	0.381 (0.056	0.270	0.491	0.950	+0.001
MI: wt, hb, age, para, **group**	0.380 (0.056)	0.270	0.491	0.940	0.000
30% missing					
Complete Case	0.380 (0.061)	0.260	0.501	0.944	0.000
MI: wt, hb, age, para	0.265 (0.065)	0.137	0.393	0.602	-0.115
MI: hb, age, para	0.263 (0.065)	0.137	0.390	0.578	-0.117
MI: hb, age, para, **group**	0.381 (0.063)	0.258	0.504	0.945	+0.001
MI: wt, hb, age, para, **group**	0.379 (0.063)	0.256	0.503	0.950	-0.001

As the proportion of missing data increased (and effective sample size decreased), the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger) at slightly different rates.

No bias was found in the effect size estimates both for the CC analyses *and* for those imputed model analyses that included study group membership in the imputation process for small to moderate missing outcome rates – and only a very small degree of bias was observed even when the missing outcome rate was as high as 30%.

Those imputation models that did not include group in the imputation process again produced (negative) bias levels of around 15% when 15% of outcomes were missing and of around 30% when the proportion of missing outcomes was as high as 30%.

Coverage was generally high for all models at all missing value levels, remaining above 0.934 (93.4%) while the proportion of missing outcomes was no greater than 5%.

However, when the proportion of missing outcomes reached 15%, coverage for the imputation models excluding group dropped to just under 88%, and when the proportion of missing outcomes reached 30%, coverage for these same models fell even further to just under 58%.

Thus, as before, when missingness in a (binary) outcome measure is MCAR, the inclusion of study group membership in the imputation models appears to greatly improve the performance of the multiple imputation procedure and provides unbiased estimates of effect size even with 30% of outcomes unrecorded. Excluding group from the imputation process, however, appears to produce marked levels of bias (tending to under-estimate the true effect size) which increase as the proportion of missing outcome values increases. These findings are consistent with those obtained when the missing outcomes are imputed as binary.

5.3.2.3 Efficacy rates 98% vs. 95%

The results of these analyses are presented in Table 5.15. These findings are virtually identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.2.3).

Table 5.15: Estimated efficacy differences, coverage and bias for different proportions of missing MCAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets *	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data	5000	0.030 (0.026)	-0.020	0.080	0.939	0.000
5% missing						
Complete Case	5000	0.030 (0.026)	-0.021	0.081	0.940	0.000
MI: wt, hb, age, para	4997	0.029 (0.026)	-0.023	0.080	0.946	-0.001
MI: hb, age, para	4994	0.028 (0.026)	-0.023	0.080	0.948	-0.002
MI: hb, age, para, **group**	4997	0.030 (0.026)	-0.021	0.082	0.940	0.000
MI: wt, hb, age, para, **group**	4997	0.030 (0.026)	-0.021	0.082	0.939	0.000
15% missing						
Complete Case	5000	0.030 (0.028)	-0.024	0.084	0.935	0.000
MI: wt, hb, age, para	4989	0.025 (0.028)	-0.029	0.080	0.961	-0.005
MI: hb, age, para	4983	0.025 (0.028)	-0.029	0.080	0.959	-0.005
MI: hb, age, para, **group**	4989	0.030 (0.028)	-0.025	0.085	0.943	0.000
MI: wt, hb, age, para, **group**	4990	0.030 (0.028)	-0.025	0.086	0.944	0.000
30% missing						
Complete Case	5000	0.030 (0.030)	-0.029	0.089	0.940	0.000
MI: wt, hb, age, para	4964	0.021 (0.030)	-0.038	0.079	0.972	-0.009
MI: hb, age, para	4969	0.021 (0.030)	-0.037	0.078	0.968	-0.009
MI: hb, age, para, **group**	4964	0.030 (0.031)	-0.032	0.091	0.947	0.000
MI: wt, hb, age, para, **group**	4971	0.030 (0.031)	-0.031	0.091	0.944	0.000

***: number of data sets for which convergent analysis was achieved.**

As found in the MAR situation reported earlier with continuous imputed outcomes, all complete case (CC) analyses converged without any problem - but a small number of imputed analyses did not converge and this problem increased as the proportion of missing outcome values increased.

In common with previous scenarios considered in which both efficacy rates were close to the boundary, the sample size was insufficient to detect an effect size as small as 0.030 (3%). All of the analyses returned statistically non-significant findings.

As the proportion of missing data increased (and effective sample size decreased), the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger) at identical rates.

The estimates of adjusted efficacy difference were unbiased for all missing value levels when CC analyses were performed, and for those imputation models which included group in the imputation process. The bias in the effect size estimates was notably greater for those imputation models that did not include group, reaching ~17% when 15% of outcomes were missing and ~33% when 30% of outcomes were missing.

Coverage was worst (0.938 compared to the set nominal level of 0.950) when the full data set was analysed. When missing outcomes were introduced, coverage ranged from

0.940 to 0.972; these higher values were again attributed to the imputation process pushing both efficacy rate estimates close to the boundary for some models.

In summary, in this scenario CC analyses and those imputation models that included group in the imputation process provided unbiased estimates of effect size with good coverage and efficiency. Only those imputation models that excluded group from the imputation process produced biased estimates of effect size (the level of bias increasing with the proportion of missing outcomes). Statistical efficiency fell at identical rates for both the CC and multiple imputation analyses as the proportion of missing outcomes increased.

5.3.2.4 Imputing MCAR continuous outcomes with binary estimates - summary

As in the MAR situation, CC analyses performed as good, and often better, than imputed model analyses, consistently yielding unbiased estimates of effect size. Some degree of efficiency was inevitably lost as the percentage of missing outcomes increased, due to the resulting decrease in effective sample size, but this was no worse than found in the imputed model analyses.

No convergence problems were detected with the CC analyses, but these were experienced when imputation models were used as an alternative method for handling missing outcomes (missing binary outcomes being replaced by continuous imputation "estimates") in the situation where both efficacy (risk) levels were close to the

parameter boundaries, due to all imputed values being allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

Even when missing binary outcomes are MCAR (i.e. when missingness is effectively wholly random), excluding study group membership from the imputation process produced biased estimates.

5.3.2.5 Missing Not At Random (MNAR) scenarios

In section 5.2.3 above, missing MNAR binary outcomes were imputed on a binary scale. This section repeats that analysis but with missing MNAR binary outcomes now imputed on a continuous scale. Only CC analyses and two multiple imputation models will be considered. The two imputation models are: (i) model with age, hb, wt, para and group; (ii) model with age, hb, wt, para. These were chosen specifically to ascertain whether the inclusion of group in the imputation models continued to play an important role in the imputation process.

5.3.3.1 Efficacy rates 85% vs. 60%

The results of these analyses are presented in Table 5.16. These findings are virtually identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.3.1).

Table 5.16: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 85% vs. 60% (RD 0.250)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.250 (0.061)	0.130	0.369	0.946	0.000
5% missing					
Complete Case	0.257 (0.063)	0.133	0.381	0.946	+0.007
MI: wt, hb, age, para	0.244 (0.064)	0.119	0.370	0.955	-0.006
MI: wt, hb, age, para, **group**	0.256 (0.064)	0.132	0.381	0.945	+0.006
15% missing					
Complete Case	0.273 (0.068)	0.140	0.407	0.932	+0.023
MI: wt, hb, age, para	0.233 (0.070)	0.096	0.371	0.972	-0.017
MI: wt, hb, age, para, **group**	0.272 (0.069)	0.137	0.407	0.933	+0.022
30% missing					
Complete Case	0.299 (0.078)	0.146	0.451	0.893	+0.049
MI: wt, hb, age, para	0.211 (0.080)	0.054	0.367	0.972	-0.039
MI: wt, hb, age, para, **group**	0.296 (0.078)	0.142	0.450	0.894	+0.046

As observed in both the MAR and MCAR cases for this scenario, as the proportion of missing data increased (and so as the effective sample size reduced), the effect size estimates from both the complete case (CC) and imputed model analyses became increasingly inefficient (i.e. the standard error of this estimate became larger).

Coverage levels were high for the imputation models that excluded group from the imputation calculations, irrespective of the proportion of outcomes that were missing.

For the complete case (CC) analyses and the imputation models that included group, coverage ranged between 93.2% and 94.6% for the 5% and 15% missing rates respectively – but when the missing rate was increased to 30% coverage fell to just below 90% which was unacceptably low.

Unlike in both the MAR and the MCAR cases, when outcomes were missing MNAR there was some degree of bias in the effect size estimates both for the CC analyses *and* for all imputed model analyses irrespective of whether group was included in the imputation process or not. Surprisingly (but confirming the findings above when the missing binary outcomes were replaced by binary estimates), for the multiple imputation analyses:

- although the differences were numerically quite small, models that included group in the imputation process exhibited greater levels of bias than models that excluded group from this process;
- models that included group produced effect size estimates that were increasingly *positively* biased as the proportion of missing outcome values increased while models that excluded group produced estimates that were progressively more *negatively* biased;
- the degree of positive bias in the multiple imputation models that used group membership was virtually identical to that in the corresponding CC analyses.

In summary, both the CC and multiple imputation models produced biased estimates of effect size. The exclusion of study group membership in the imputation models maintained coverage levels but produced negatively biased estimates of effect size. Including study group membership produced virtually identical results to the CC analysis, with coverage markedly reduced and effect size consistently over-estimated.

5.3.3.2 Efficacy rates 98% vs. 60%

The results of these analyses are presented in Table 5.17. These findings are nearly identical to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.3.2).

Table 5.17: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 60% (RD 0.380)

Model	RD (SE)	95% CI*		Coverage	Bias
		LL	UL		
Full data	0.380 (0.051)	0.279	0.480	0.950	0.000
5% missing					
Complete Case	0.394 (0.053)	0.291	0.497	0.948	+0.014
MI: wt, hb, age, para	0.374 (0.054)	0.267	0.481	0.955	-0.006
MI: wt, hb, age, para, **group**	0.394 (0.053)	0.289	0.498	0.948	+0.014
15% missing					
Complete Case	0.427 (0.056)	0.317	0.537	0.868	+0.047
MI: wt, hb, age, para	0.361 (0.062)	0.239	0.482	0.974	-0.019
MI: wt, hb, age, para, **group**	0.427 (0.059)	0.311	0.543	0.884	+0.047
30% missing					
Complete Case	0.486 (0.062)	0.366	0.607	0.584	+0.106
MI: wt, hb, age, para	0.333 (0.073)	0.189	0.477	0.967	-0.047
MI: wt, hb, age, para, **group**	0.486 (0.069)	0.350	0.621	0.667	+0.106

As for the previous scenarios, as the proportion of missing data increased, the effect size estimates from both the complete case (CC) and imputed model analyses became progressively inefficient (i.e. the standard error of this estimate increased).

Irrespective of whether or not group was included in the imputation calculations, the imputation models as well as the complete case analyses returned biased effect size

estimates. The imputed models that included group produced more biased estimates than those that excluded group. The effect size estimates from those imputation models that included group in their calculations were positively biased for all levels of missing data, whereas the estimates from imputation models that excluded group were negatively biased.

For all missing outcome levels, imputation models that excluded group produced estimates of efficacy difference that were very efficient (coverage rates were actually higher than the expected 95%). Coverage rates for both the CC analyses and the imputation models that included group in the imputation process were reasonable when only 5% of outcomes were missing, but fell to 86.8% and 88.4% respectively when the missing outcome proportion was 15%; when this proportion was increased further to 30%, coverage was reduced to the unacceptably low levels of 58.4% and 66.7% respectively.

In summary, the CC and multiple imputation analyses were all biased in this situation. Including study group membership in the multiple imputation models exacerbated rather than improved performance, and this became more marked as the proportion of missing outcomes was increased. Excluding group membership from the computation of the multiple imputation models maintained coverage levels but produced negatively biased estimates of effect size, while including study group produced virtually identical results to the CC analysis, with coverage now markedly reduced and effect size consistently over-estimated.

5.3.3.3 Efficacy rates 98% vs. 95%

The results of these analyses are presented in Table 5.18. These findings are similar to those obtained with outcome estimated as a binary rather than as a continuous variable (section 5.2.3.3).

Table 5.18: Estimated efficacy differences, coverage and bias for different proportions of missing MNAR outcomes-continuous imputed outcomes (averaged over 5000 imputed data sets): efficacy rates 98% vs. 95% (RD 0.030)

Model	No. of datasets *	RD (SE)	95% CI*		Coverage	Bias
			LL	UL		
Full data	5000	0.030 (0.026)	-0.020	0.080	0.939	0.000
5% missing						
Complete Case	5000	0.032 (0.028)	-0.022	0.086	0.949	+0.002
MI: wt, hb, age, para	4994	0.030 (0.028)	-0.023	0.087	0.957	0.000
MI: wt, hb, age, para, **group**	4995	0.032 (0.028)	-0.023	0.087	0.947	+0.002
15% missing						
Complete Case	5000	0.039 (0.033)	-0.026	0.103	0.953	+0.009
MI: wt, hb, age, para	4996	0.029 (0.033)	-0.036	0.095	0.979	-0.001
MI: wt, hb, age, para, **group**	4994	0.039 (0.034)	-0.029	0.106	0.944	+0.009
30% missing						
Complete Case	5000	0.052 (0.045)	-0.037	0.141	0.939	+0.022
MI: wt, hb, age, para	4994	0.028 (0.043)	-0.055	0.112	0.991	-0.002
MI: wt, hb, age, para, **group**	4994	0.052 (0.048)	-0.042	0.146	0.925	+0.022

***: number of data sets for which convergent analysis was achieved.**

As in the similar effect size difference situations reported above, all complete case (CC) analyses converged but a small number of imputed analyses failed to converge, with this problem increasing as the proportion of missing outcome values rose.

The sample size proved insufficient to detect an effect size as small as 0.030 (3%); all of the analyses returned statistically non-significant findings.

Increasing the proportion of missing data (and hence decreasing the effective sample size) increased the standard errors of the effect size estimates and decreased statistical efficiency in all analyses.

The imputed model analyses that excluded group from the imputation computations produced effect size estimates with little or no bias. However, the CC analyses and the imputed models that included group produced positively biased estimates, and degree of bias increased as the proportion of missing outcomes rose. When 30% of outcomes were missing, both of these models produced identical bias levels of 0.022; although numerically small, as the true efficacy difference was only 0.030, this represents a non-ignorable bias level of 73%. Coverage was at least 92.5% for all scenarios and all levels of missing data.

In summary, complete case analyses and multiple imputation models that included group in the imputation procedure produced biased effect size estimates (bias reaching

almost 75% when the proportion of missing outcomes was raised to 30%), whereas multiple imputation models that excluded group from the calculations were virtually unbiased for all missing outcome proportions.

5.3.3.4 Imputing MNAR binary outcomes with continuous estimates - summary

Unlike in the MAR and MCAR situations, both CC analyses and the imputed model analyses (irrespective of whether group is included or not) tended to produce biased estimates of effect size, with (in general terms) the level of bias increasing as the proportion of missing outcomes rose.

The most striking finding was that the exclusion of group from the calculations improved the performance of multiple imputation models, although the estimates produced remain biased. The estimates obtained from the complete case analyses and from those multiple imputation models that included group in the imputation process were consistently biased away from the null hypothesis (i.e. over-estimated effect size) while excluding group from the imputation analyses produced bias toward the null hypothesis (i.e. under-estimated effect size). This latter finding was consistent with the findings from all other scenarios considered in this dissertation in which group was excluded from the imputation models, but the former finding was unexpected.

Inevitably, some degree of efficiency was lost as the percentage of missing outcomes was increased in both the complete case analysis and the imputed model analyses

(irrespective of whether group was included or not). Coverage was at least 0.947 (94.7%) in the imputed models that excluded group for all scenarios considered.

No convergence problems were detected with the CC analyses, but some non-convergence was experienced when imputation models were used as an alternative method for handling missing outcomes, despite the fact that the missing binary outcomes were now being replaced by continuous imputation "estimates". Convergence problem were greatest when both efficacy (risk) levels were close to the parameter boundaries, suggesting that even imputing binary outcomes on a continuous scale, it is possible for all imputed values to be allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

5.3.3.5 Mathematical explanation of bias findings in the MNAR situation

The relationship found in the MNAR situation using multiple imputation methods between the direction of the bias in the effect size estimates and the use of group in the imputation calculations was unexpected. There is a simple mathematical explanation for this finding, however, based on the fact that, in the MNAR situation modeled, cases that had a treatment success were given a larger probability of having a missing outcome than the treatment failure cases. This is described in table 5.19 below using a simple simulated example with 1000 patients in each of groups A and B.

Suppose that:

- the efficacy rates in groups A and B are actually 80% and 60% respectively;
- the probability of the outcome being missing is 20% if the outcome is positive and 10% if the outcome is negative in both treatment groups.

Table 5.19: The effect size estimate <u>In the *absence* of MNAR</u>:

Group:		**A**	**B**
Sample size		1000	1000
Probability of:	positive outcome occurring	0.80	0.60
	positive outcome missing	0	0
	positive outcome not missing	1	1
	negative outcome occurring	0.20	0.40
	negative outcome missing	0	0
	negative outcome not missing	1	1
Expected number of:	positive outcomes	1000 * 0.80 * 1 = 800	1000 * 0.60 * 1 = 600
	negative outcomes	1000 * 0.20 * 1 = 200	1000 * 0.40 * 1 = 400
Estimated efficacy:		800 / (800 + 200) = 80.0%	600 / (600 + 400) = 60.0%
Estimated effect size:		**80.0 – 60.0 = 20.0%**	

When there are no missing outcomes, the effect size estimate will be unbiased.

In the presence of MNAR data, the mathematical explanation for the CC finding is presented in table 5.20 below. As shown in the table below, in a MNAR situation, the effect size estimate from a complete case (CC) analysis will be positively biased.

Table 5.20: The effect size estimate <u>In the *presence* of MNAR</u>: *CC analysis*

Group:		**A**	**B**
Sample size		1000	1000
Probability of:	positive outcome occurring	0.80	0.60
	positive outcome missing	0.20	0.20
	positive outcome not missing	0.80	0.80
	negative outcome occurring	0.20	0.40
	negative outcome missing	0.10	0.10
	negative outcome not missing	0.90	0.90
Expected number of:	positive outcomes	1000 * 0.80 * 0.80 = 640	1000 * 0.60 * 0.80 = 480
	negative outcomes	1000 * 0.20 * 0.90 = 180	1000 * 0.40 * 0.90 = 360
	missing outcomes	1000 - (640 + 180) = 180	1000 - (480 + 360) = 160
Estimated efficacy:		640 / (640 + 180) = 78.0%	480 / (480 + 360) = 57.1%
Estimated effect size:		**78.0 – 57.1 = 20.9%**	

Multiple imputation analysis with group excluded from imputations

In this situation, the multiple imputation process does not distinguish between the two groups when estimating whether a missing outcome should be replaced by a positive or negative outcome. Missing outcomes in both groups are thus replaced using the pooled efficacy rate for the two groups combined. The mathematical explanation is presented in table 5.21 below.

Table 5.21: The effect size estimate In the *presence* of MNAR: Excluding group in imputation models

Group:		**A**	**B**
Sample size		1000	1000
Probability of:	positive outcome occurring	0.80	0.60
	positive outcome missing	0.20	0.20
	positive outcome not missing	0.80	0.80
	negative outcome occurring	0.20	0.40
	negative outcome missing	0.10	0.10
	negative outcome not missing	0.90	0.90
Before imputation:			
Expected number of:	positive outcomes	1000 * 0.80 * 0.80 = 640	1000 * 0.60 * 0.80 = 480
	negative outcomes	1000 * 0.20 * 0.90 = 180	1000 * 0.40 * 0.90 = 360
	missing outcomes	1000 - (640 + 180) = 180	1000 - (480 + 360) = 160
Estimated efficacy:	per group	640 / (640 + 180) = 78.0%	480 / (480 + 360) = 57.1%
	Overall	(640 + 480) / (640 + 180 + 480 + 360) = 67.5%	
After imputation:			
Estimated number of:	positive outcomes	640 + (180 * 0.675) = 761.5	480 + (160 * 0.675) = 588
	negative outcomes	180 + (180 * 0.325) = 238.5	360 + (160 * 0.325) = 412
Estimated efficacy:		761.5 / (761.5 + 238.5) = 76.15%	588 / (588 + 412) = 58.80%

Estimated effect size:	**76.15 - 58.80 = 17.35%**

In a MNAR situation, the effect size estimate from a multiple imputation analysis in which group membership is *excluded* from the imputation process will be negatively biased.

Multiple imputation analysis with group included in imputations

In this situation, the multiple imputation process does distinguish between the two groups when estimating whether a missing outcome should be replaced by a positive or negative outcome. Each missing outcomes is thus replaced using the efficacy rate for the group in which that missing outcome occurred. The mathematical explanation is detailed in table 5.22 below.

Table 5.22: The effect size estimate In the *presence* of MNAR: including group in imputation models

Group:		**A**	**B**
Sample size		1000	1000
Probability of:	positive outcome occurring	0.80	0.60
	positive outcome missing	0.20	0.20
	positive outcome not missing	0.80	0.80
	negative outcome occurring	0.20	0.40
	negative outcome missing	0.10	0.10
	negative outcome not missing	0.90	0.90
Before imputation:			
Expected number of:	positive outcomes	1000 * 0.80 * 0.80 = 640	1000 * 0.60 * 0.80 = 480
	negative outcomes	1000 * 0.20 * 0.90 = 180	1000 * 0.40 * 0.90 = 360
	missing outcomes	1000 - (640 + 180) = 180	1000 - (480 + 360) = 160
Estimated efficacy:	per group	640 / (640 + 180) = 78.0%	480 / (480 + 360) = 57.1%
	overall	(640 + 480) / (640 + 180 + 480 + 360) = 67.5%	
After imputation:			
Estimated number of:	positive outcomes	640 + (180 * 0.780) = 780.4	480 + (160 * 0.571) = 571.4
	negative outcomes	180 + (180 * 0.220) = 219.6	360 + (160 * 0.429) = 428.6
Estimated efficacy:		780.4 / (780.4 + 219.6) = 78.04%	571.4 / (571.4 + 428.6) = 57.14%

Estimated effect size:	**78.04 – 57.14 = 20.90%**

In a MNAR situation, the effect size estimate from a multiple imputation analysis in which group membership is *included* in the imputation process will be positively biased – and (theoretically) the extent of this bias will be the same as that in a CC analysis.

Chapter 6 : Discussion and conclusions

6.1 The Binomial regression model, Copy method and Cheung's OLS method

The binomial regression model with an identity link function is prone to convergence problems in software that uses MLE. From the simulation findings in this study, non-convergence rates were found to increase as at least one of the efficacy rates moved towards a boundary value irrespective of the number of covariates included in the substantive model. For all scenarios examined, convergence was poor when the efficacy rate in either group was 90% or greater. This finding on the boundary efficacy rates is consistent with previous publications (Wacholder 1986). Treatments are now becoming very effective in malaria studies, so it is common to have at least one efficacy rate (usually for the intervention group) near a boundary value (usually 100%). It has been suggested (Borrmann et al. 2008) that anti-malarial drugs should be recommended for global use only when efficacy is at least 90%. Thus, in malaria studies attempting to report adjusted efficacy differences close to 100%, the binomial regression model is likely to be highly susceptible to model failure (non-convergence) severely limiting its ability to provide unbiased estimates (or, indeed, any estimates at all) in this situation.

It was interesting to find that this convergence problem worsened as the number of covariates in the model was increased. Models with just one covariate tended to have less convergence problems than models with two covariates, and in turn these tended have fewer convergence problems than models with three covariates, and so on. This

means that the more covariates that are included in the model, the higher the probability that the binomial regression model will fail to converge and produce a reliable estimate of effect size. In practice, however, many potential confounders usually exists that need to be adjusted for in regression models in order to obtain an accurate estimate of true treatment success (failure) that is independent of other associated factors. Clearly, therefore, the use of the binomial regression model with an identity link function to obtain an accurate estimate of effect size while controlling for potential confounding variables is probably limited to situations in which efficacy in all treatment groups being compared is modest or poor (i.e. ≤90%).

The degree of correlation between confounders / covariates also appears to be associated with the risk that the binomial regression model will fail. In the simulations conducted for this dissertation, the percentage of datasets that converged improved when the correlations between the covariates were removed in all of the models considered. The improvement in convergence was most remarkable in models with a large number of covariates. Covariates are supposed to be statistically independent and hence uncorrelated, but in practice there is always some degree of correlation between covariates. The findings of this study appear to indicate that the application of the binomial regression model to obtain adjusted risk (efficacy) differences is probably restricted to those situations in which there is truly little or (ideally) no correlation between the covariates being adjusted for.

The COPY method is known to reduce non-convergence problems when attempting to estimate a risk ratio using a log-binomial regression model (Savu et al. 2010). Indeed, for a log-binomial regression model, the COPY method provides an effective method of estimating prevalence ratios, as long as the starting values are appropriate (Savu et al. 2010). Unfortunately, however, although the COPY method is simple to apply, and of course intuitively striking, it appears not to be an appropriate approach for dealing with the problem of non-convergence when trying to obtain unbiased estimates of the adjusted efficacy (risk) differences using a binomial regression model with identity link function. With the binomial regression model, the number of copies required to minimize the risk of a model failing to converge was found to coincide with the number of copies that gave the most biased estimates of the true efficacy difference. Counter-intuitively, increasing the number of copies made both the risk of non-convergence and bias in the efficacy difference estimates worse. So, the COPY method is probably best suited to risk ratio modeling using the log-binomial regression model.

The main challenge in practice is that researchers generate only one set of data and if that particular dataset does not converge in the available software package(s), then alternative approaches are immediately needed. The most useful method would be one that has 100% convergence and produces unbiased estimates of the risk differences. While the COPY method used in conjunction with the binomial regression model and an identity link function clearly does not always achieve this ideal, it should not be entirely dismissed as a possible analysis option. Given that, mathematically, binomial regression is the best method for estimating effect size differences if the technique

works, it should be attempted as the first choice analysis. If, however, the model fails to converge, the COPY method should be attempted, starting with a relatively small number of copies (e.g. 5). The number of copies can be increased to a maximum of 10; if the model is still failing to converge, there is little point in adding further copies as this is extremely unlikely to achieve convergence. If convergence is achieved with 10 or less copies, the effect size estimate is likely to be biased to some degree and so a secondary CC analysis may be required to estimate the degree of this bias. This policy should initially be implemented with no covariate adjustment; if convergence problems are experienced even with the addition of the COPY method, it is likely that these problems will increase when covariate adjustment is attempted, so it might be sensible to look for an alternative analysis approach rather than to attempt covariate adjustment with the binomial regression model.

A computationally very attractive alternative method is Cheung's modified OLS with Huber-White robust standard errors. This is a simple regression method that is used for the estimation of the risk differences and can be applied using any standard statistical software that performs ordinary least-squares regression and has options for obtaining Huber-White standard errors. This is probably one of the most attractive features of this method. Furthermore, the method offers statistical advantages over the standard binomial regression model (with or without the COPY method) because convergence is more (but not totally) guaranteed. In addition, the method yields unbiased estimates of adjusted risk differences, and valid standard errors can be obtained using the Huber-White formulae. On the downside, however, unlike the binomial regression model

where exact confidence intervals are obtained for the effect size estimate when the model is convergent, Cheung's modified OLS only provides model based (approximate) estimates of the confidence intervals.

Thus, if covariate adjusted efficacy (risk) differences are of epidemiological interest, the standard binomial regression model with identity link function, even when modified using the COPY method, should not be the sole method for estimating effect size differences listed in the statistical analysis plan during the study protocol development. The Cheung's modified OLS method must be specified as an alternative approach to be used should the binomial model fail.

In summary, therefore:

- the binomial regression model with an identity link function faces an increasing risk of convergence challenges especially when at least one of the efficacy rates moves towards a boundary;
- the risk of these convergence problems increases as number of covariates being adjusted for increases and also as the correlation between these covariates increases
- although the COPY method is simple to apply, is intuitively attractive, and is an effective approach for handling the problem of non-convergence in other situations (e.g. when estimating odds ratios or risk ratios), it does not appear to be as effective when used with the binomial model to estimate risk differences;

- even when the COPY method does work (i.e. achieves model convergence), it does not produce unbiased estimates of the adjusted risk difference;
- Cheung's modified OLS with Huber-White robust standard errors is a sensible alternative method when the binomial regression model does not converge, having the appealing features of being simple to apply in all standard software and assuring convergence to a reliable effect difference estimate.

6.2 Comparison of methods of handling missing data

6.2.1 Imputing MAR binary outcomes

For data that was MAR, the performance of Complete Case (CC) analyses was found to be as good as, and often better than, using imputed model analyses, consistently producing unbiased estimates of efficacy difference. Under this missing mechanism, the performance of CC analyses was found to unaffected by whether the imputed values for the missing *binary* outcomes were estimated on a binary or on a continuous scale. This finding is consistent with a recent publication by Groenwold et al (2011) who examined missing binary outcomes in a randomized trial setting using ***odds ratios*** as the summary statistic of interest for estimating the relative effects of two or more treatments and also found that the performance of the CC analysis method was either the same as, or better than the MI approach. However, since ***odds ratio modeling*** and risk difference modeling use different mathematical algorithms, it cannot be assumed that the results from an odds ratio model can be extrapolated to a risk difference model. So in this study, the performance of the CC analysis and MI method were investigated using simulation

methods with risk difference now as the outcome measure of interest instead of odds ratios.

Under the MAR situation, there was some predictable loss of statistical efficiency in the CC analyses as the percentage of missing outcomes increased due to the diminishing effective sample size, as has previously been reported (Desai et al. 2011). However, the loss in efficiency was the same as, and often better than (i.e. numerically less than), that observed in the multiple imputation analyses. This is surprising and unexpected. Theoretically MI methods are expected to yield correct standard errors (Donders et al. 2006), for two reasons. Firstly, sample size is maintained; secondly uncertainty in the imputed values is fully taken into account.

A plausible explanation for this unexpected finding is that, although sample size is maintained when using MI procedures, these same procedures also increase the variability in the outcome values (irrespective of whether binary or continuous outcome values are being imputed) that acts to inflate the standard error of the effect size estimate. This increase in variability is likely to result from the random component that is added to the missing outcome values during the imputation process (Rubin 1987, Rubin 1996, Collins et al. 2001, Little 2002, Groenwold et al. 2011). For this study, ten imputations were used in each multiple imputation procedure, based on published recommendations (Rubin 1987, Schafer 1997) that using between 3 and 10 imputations

should be sufficient to obtain valid estimates of the effect size parameters. However, Bodner (2008) showed that using between 3 and 10 imputations may result in the important parameters of interest suffering from huge imputation variability. This study appears to confirm this finding, but further research is needed into this important influence on the precision of risk / efficacy difference estimates.

No convergence problems were experienced with the CC analyses under the MAR condition. However, convergence problems did occur in the MAR condition when imputation models were used to deal with missing outcomes where both efficacy rates were close to the parameter boundaries. This, evidently, was due to all imputed values being allocated to the same outcome value (which can easily happen when imputing only few values) resulting in zero standard errors for the effect size estimate. This phenomenon is often referred to as "perfect prediction" (Royston and White 2011).

Although this phenomenon this occurred regardless of whether the imputed outcomes were on a binary or on a continuous scale, perhaps predictably the problem was relatively worse in those situations where the outcomes were imputed on a binary rather than on a continuous scale. When imputing outcomes on a binary scale there are only two possible values that the outcome being imputed can take - either zero or one. So for example, if either or both efficacies are high (i.e. $\geq$95%) and there are relatively few missing outcome values, the probability of all of these missing outcomes being

estimated as the same value (in this case 1) is likely to be high. When imputing (binary) outcomes on a continuous, theoretically each imputed missing outcome can take infinitely many values between 0 and 1; even so, when either or both efficacies are high (i.e. ≥95%) and there are relatively few missing outcome values, the probability of all of these missing outcomes being estimated as the same value (in this case 1) is still likely to be high.

This problem of "perfect prediction" can easily be resolved in Stata by using the command option "augment", which causes an augmented regression to be performed (Royston and White 2011). In the simulations reported from this study, the augment option was used in all situations where the efficacy rates were 95% or 98% for the two comparator groups. While the use of this option had a big impact on the overall levels of perfect prediction that occurred, it was observed that the augment option only helped to reduce the problem but did not completely eliminate it.

Another striking finding under the MAR condition was that the inclusion of treatment group membership in the imputation process (i.e. the use of this variable in the calculations to estimate the missing outcome values) played a crucial role in improving the performance of the imputation process. If missingness in a binary outcome is MAR and related to study group membership, excluding this variable from the imputation process was found to produce biased estimates of the adjusted efficacy (risk) difference,

whereas if group was included in the process, the estimate was unbiased. However, if missingness is related *also* to other factors or covariates, the absence of these additional factors or 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* even in the situation where one of the effect sizes was close to a boundary value. Clearly, therefore, if missing outcomes are MAR and related to several known factors *including treatment group membership*, including group in the imputation process is paramount over all other factors related to missingness.

The MAR assumption is often a plausible assumption in practice and there are many clinical trial situations in which it is likely that missing outcomes will be related to other observed variables (Schafer and Graham 2002, Kenward and Carpenter 2007). For example, it is plausible that the probability of a participant dropping out of clinical trial could be associated with the treatment to which the participant was allocated (Altman 2009). The above findings under the MAR assumption thus occupy a very important role in many randomized trials, especially those whose measure of effect is a risk difference.

In summary, when some values for a binary outcome variable are MAR (missing at random) but all observations are recorded for all other baseline variables, a CC analysis produces adjusted estimates that are unbiased with no obvious risk of convergence problems so would seem to be the analysis method of choice. This view would appear

to be further supported by the finding in this study that, under the MAR condition, CC often performs as well as or better than multiple imputation methods over a wide range of efficacy rates. In addition, CC analyses are easy to implement since this is often the default method in statistical software packages.

Unfortunately, however, the CC analysis method is not consistent with the intention to treat (ITT) principle, which is the standard approach for the analysis of a RCT. The ITT approach requires that all subjects that were randomized should be included in the analysis according to their randomization allocation. So, while CC is preferable to MI methods on the basis of both performance and ease of application, the ITT principle probably over-rides this and so MI should be the analysis method of first choice, with a CC analysis performed as secondary (confirmatory / sensitivity) analysis strategy.

6.2.2 Imputing MCAR binary outcomes

Although it is often difficult for study outcome data to be MCAR, this condition is often plausible in malaria efficacy studies where missing outcome is due to for example: indeterminate outcomes resulting from the presence of mixed genotypes in the post treatment samples; a test tube containing sample being broken before sample processing; sample processing failure in a laboratory equipment; samples becoming haemolysed and/or clotted. Such events usually happen completely at random and making the MCAR assumption valid.

As in the MAR situation, CC analyses were found to perform as well as, and often better than, multiple imputed methods, consistently producing unbiased estimates of effect size. This is again expected from theory under-pinning CC analyses - the subjects remaining in a CC analysis constitute a random sample of the target population (Rubin 1987, Little 2002, Schafer and Graham 2002, Carpenter and Kenward 2006, Carpenter and Kenward 2007).

There was to some extent, again, some degree of efficiency inevitably lost in the CC analyses as the percentage of missing binary outcomes increased. As explained for MAR situation above, this was due to the resulting decrease in effective sample size, but the losses in efficiency observed were no worse than those found in the multiple imputation analyses. Again, a plausible explanation is that, although sample size is maintained in MI procedures, there is an increased variability in the imputed outcome values that tends to inflate the standard errors. This increase in the variability of the outcome values in MI most probably results from the random component that is added to the estimates of the missing outcome values during the imputation process (Rubin 1987, Rubin 1996, Collins et al. 2001, Little 2002, Groenwold et al. 2011).

As in the MAR situation, no convergence problems were experienced with the CC analyses. Convergence problems were experienced, however, when imputation models were used as an alternative method for handling missing outcomes in the situation where

both efficacy (risk) levels were close to the parameter boundaries. These occurred irrespective of whether the missing binary outcomes were replaced by binary or continuous imputation "estimates") and, again as in the MAR situation, was attributed to all imputed values being allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

Finally, in yet another similar finding to the MAR condition, when missing binary outcomes were MCAR (i.e. when missingness was effectively wholly random), excluding study group membership from the imputation process was found to produce biased estimates.

In summary, therefore, as for the MAR situation, when some values are unrecorded for a binary outcome variable but recorded for all other variables, CC analysis would appear to be the method of choice for handling missing binary outcomes. Crucially, unlike the imputation models, the CC the efficacy (risk) difference regression models will converge. So, as a general recommendation based on the simulation studies, CC would appear to be preferable to MI methods as the latter offers no statistical advantages. However, the intention to treat principle (ITT), the standard for analysing and reporting RCTs, probably over-rides this and so MI should be analysis of first choice, with a CC analysis performed as secondary analysis strategy for verification and sensitivity purposes. The CC should be used for the "per protocol analyses" as MI has no place in per "protocol analyses". If the MI and CC results disagree, the results of both CC and MI should be reported and possible reasons for the discrepancy discussed.

Needless to say, in both the MAR and MCAR situations, all of the above problems could be either avoided or, at worst, minimised by maximising the collection of the outcome measures.

6.2.3 Imputing MNAR binary outcomes

The possibility of having missing outcomes that are MNAR cannot be ignored in practice. Indeed, it is possible that missing outcomes can be directly related to the outcome itself. For example, individuals on an inferior treatment may not find any benefit from the treatment and so may be more likely to have a treatment failure or to drop-out than those individuals on the superior treatment, in which case (some) missing outcomes will be MNAR.

Particular attention was directed in this study to the situation in which MNAR was linked to group membership. The simulation exercises were set up deliberately so that missingness in the binary outcome measure was related to treatment group membership; patients in the more effective treatment group were given a higher probability in the simulation calculations of having a missing outcome than those in the other, less effective, treatment group. In this situation, the findings differed considerably from those obtained under both the MAR and MCAR conditions. Under the MNAR condition, both the CC analyses and the multiple imputation analyses were biased, in the latter case irrespective of whether or not group was included in the imputation process.

This finding is essentially in line with published theory, which states that a wrong model is being used in both the CC and MI analyses (Rubin 1976, Rubin 1987, Little 2002, Liublinska and Rubin 2012). The really striking finding, however, was that although missingness was related to treatment group membership, including group membership in the imputation calculations now *increased* the degree of bias in the effect size estimates compared to excluding group from the imputation process. An examination of the mathematical implications of MNAR indicated that, when the aim of the analysis is to asses ***group*** efficacy (risk) difference, if the MNAR condition is linked directly to group membership, in both the CC and MI analysesthe MNAR missingness will in itself bias the effect size estimate – and that the inclusion of the ***group*** variable in a MI analysis actually reinforces and exaggerates the level of bias in the effect size estimate. It is likely that this effect is less pronounced in a MI analysis for variables other than group membership, that is, if any other variables are related to missingness and are included in the MI calculations, this will also increase the bias in the effect size estimate, but to a lesser degree than will occur with group.

Both the CC and MI analyses that include group in the imputation process produced "positive" bias, in that the effect size estimate tended to be consistently pushed away from the null hypothesis (i.e. the effect size tended to be consistently over-estimated), whereas those MI analyses that excluded group tended to push the bias towards the null hypothesis (i.e. the effect size now tended to be consistently under-estimated).

In common with both the MAR and MCAR condition, no convergence problems were detected with the CC analyses under MNAR. Convergence issues did arise, however, when multiple imputation methods, irrespective of whether the missing binary outcomes were replaced by binary or continuous imputation "estimates", and this problem was particularly acute in the situation where both efficacy (risk) levels were close to the parameter boundaries, due to all imputed values being allocated to the same outcome value (resulting in zero standard errors for the effect size estimate).

In summary, under the MNAR condition for missing binary outcomes when missingness is (either wholly or in part) related to treatment group membership, both CC and MI analyses produce biased estimates of effect size (effect difference). Furthermore, the inclusion of group in a multiple imputation analysis will tend to exaggerate bias away from the Null hypothesis (i.e. will tend to over-estimate the true effect size); exactly the same impact on bias will occur if a CC analysis is carried out in this situation. Excluding group from the calculations in a MI analysis produces less bias, but now the bias is towards the Null hypothesis (i.e. the true effect size will tend to be under-estimated).

Thus, multiple imputation methods may have an advantage over CC methods when outcomes are MNAR in the sense that, if the variables that cause the non-ignorable (MNAR) missing outcomes are known, these variables can be included in the imputation calculations. It was not possible in this study to examine the effects on bias

when variables other than group membership that influence missingness are included in the imputation calculations, and further research is needed on this important issue – but intuitively it is possible that such variables will, if included in the imputations, reduce bias (rather than increase bias as happens when group is included). For this reason, MI methods play a very important role in performing sensitivity analyses when data are MNAR (Carpenter et al. 2007, Kenward and Carpenter 2007). For longitudinal studies in which outcome is measured on several occasions, Diggle and Kenward 1994 proposed a parametric model for analyzing such designs when there are non-ignorable dropouts (Diggle and Kenward 1994).

6.3 Consort statement and WHO recommendations on handling missing data in RCTs

The Consort statement states that the statistical analysis strategy for a RCT should be clearly stated in advance(Altman et al. 2001), and that this strategy should be based on either the ITT (intention to treat) principle or the "on-treatment" principle (commonly known as per protocol (PP) principle).

Under the intention to treat principle, all participants recruited in the study are included in the statistical analyses within the group to which they were originally randomized. The challenge with this is that some of the participants may not have outcomes. Although the Consort statement acknowledges that the ITT approach has the advantage of reducing any bias resulting from systematic loss of participants (e.g. Lee (1991) and

Lachin (2000)), it remains unclear how to include those individuals with some missing outcomes. The findings of this study probably provide some guidance on this issue for the ITT strategy. MI may be a useful approach for handling missing outcomes in order to be consistent with the ITT principle. However, for the MI approach to be valid, data needs to be MAR or MCAR. For MNAR data, the use of MI should be followed by sensitivity analyses.(The Consort statement stipulates that the ITT approach is not appropriate for assessing adverse events so advises performing additional sensitivity analyses in this situation also).

The Consort Guidelines offers the authors of RCT reports strong advice on two important issues:

- What constitutes ITT in their analyses must be clearly defined; the original Consort document noted that, out of 249 articles that were reviewed in 1997, only 2% clearly stated that all participants who were randomized were analyzed according to the group they were randomized to.
- Removing participants from the statistical analysis process could, irrespective of the reasons for exclusion, lead to invalid inferences.

The 2003 WHO report on the assessment and monitoring of antimalarial drug efficacy for the treatment of uncomplicated falciparum malaria recommends the application of the intention-to-treat (ITT) principle for the primary statistical analysis (apart from when survival analyses are being used) followed by secondary per protocol analyses. The

results from the secondary per protocol analyses allow comparisons with any available historical results.

Clearly, where per protocol analyses are performed, MI does not have any place so the CC analysis approach should be employed. This meshes well with the findings of this thesis that CC analyses are often better than MI provided the missing binary outcomes are either MAR or MCAR. It is emphasized again that this discussion must be considered in the context of a RCT design in which a binary outcome is collected at just one time point of interest.

6.4 The effect of missing values and the validity of the missing data simulation findings

A number of important assumptions were employed in the simulation studies reported in this thesis especially relating to the methods used in the handling of missing data, ranging from study design, the outcome measure of interests, efficacy levels, missingness being limited to the outcome variable, missingness levels and missingness mechanisms. Each of these are now considered in turn.

6.4.1 Study design

The simulated data used in this study mimicked a real randomized cohort study for a special case in which the outcome of interest was measured only once-usually at the end

of the study. This is often the case in randomized controlled trials of malaria efficacy. In such studies, the status of study participants at day 28 following the onset of treatment is often of particular interest. Thus, although measurements are taken at several time points during follow up, the "headline" statistical analysis of the primary outcome measure is often based on the outcome data from just this one time point.

It is emphasized at this point that the findings on the methods of handling missing data reported in this dissertation are limited to just those studies that are randomized and in which the outcome at only one time point is of interest. The results should not be extrapolated to longitudinal study designs in which measurements taken at several time points of interest. In such studies, the levels of missing data are usually high and methods such as complete case analysis may no longer be the appropriate for dealing with missing outcome data.

6.4.2 Outcome measure

The simulation studies reported in this dissertation are based on a binary outcome and the measure of effect of interest was risk difference. It is stressed, therefore, that the findings are limited to scenarios where the outcome of interest is risk difference. It does not matter, however, whether missing values are imputed on a binary or on a continuous scale.

6.4.3 Efficacy levels

The simulations covered a wide range of efficacy levels, including close to the 100% boundary, so the findings in this dissertation apply over the full range of efficacy values.

6.4.4 Missingness, missingness levels and missingness mechanisms

The simulations considered a specific scenario in which missing data was confined to the outcome variable. This is often the case in randomized studies of malaria efficacy. Most of the baseline information data is collected during screening as these often form part of the inclusion and exclusion criteria. Therefore the findings of these simulations should not be extended to observational studies where there may be high levels of missing data in the explanatory variables.

Missing levels for the outcome measure of up to 30% were considered; it remains unknown what the findings may be for missing levels that go beyond 30%. However, missing levels greater than 30% in the outcome measure during a randomized trial probably indicates an important flaw in the design of the study that may have both ethical and statistical implications (e.g. is the power of the study may be drastically reduced).

In terms of the process that may be creating missing outcomes, all three of the categories of missing data mechanisms described by Rubin (1976) were considered in

the simulations. Thus, the findings in this dissertation hold for these specific missing data mechanisms.

6.5 Bias towards the null observed when wrong models are used (MAR and MCAR)

In situations where both group membership and the value of the (binary) outcome measure are linked to missingness, it has been shown in the simulations in this thesis that omission of the group variable from the imputation process tends to bias the estimate of effect size towards the null. The reason for this is that, if group is omitted from the imputation process, the imputed outcome values will not be associated with the outcome while observed values will. The overall association between the group and outcome will be reduced producing bias that goes toward the null hypothesis. This will increase type II error thereby reducing the power of the study to detect the difference in effect in the two study groups. A Complete Case analysis maintains the association that exists between the outcome and the group variable that is linked to both outcome and missingness. Consequently the results from complete case analyses are often better than those from MI when data are MAR or MCAR.

6.6 Perfect prediction in MI procedures

It was observed in the simulation findings that when both efficacy rates were close to 100% (boundary value), some of the models failed to produce an output (i.e. failed to

produce an estimate of effect size) when MI was used – and this happened irrespective of whether the imputed values were considered to be binary or continuous.

The explanation for this is probably quite simple and straightforward. Perfect prediction (i.e. an estimate of effect size with no error due to all participants having the same value for the outcome measure) can arise in any Generalized Linear Model (GLM) that has a categorical outcome(White et al. 2010). In the case of the simulations carried out for this thesis, it is likely that the reason for perfect prediction in a MI analysis was that all the imputed values took the same value across all participants, resulting in zero variance estimates and making the calculation of degrees of freedom impossible as division by zero is impossible.

The degrees of freedom v for the calculation of confidence intervals and statistical tests from MI, is given as:

$v = (p-1)\left[1+\frac{\overline{\vartheta}}{(1+p^{-1})\xi}\right]^2$, where ξ is the between imputation variance as given by Rubin (1976). So when ξ is zero, there is likely to be the perfect prediction problem.

White et al (2010) additionally suggest that the problem can arise with standard errors that are extremely large, as these usually reflect the approximately flat nature of the likelihood.

In order to overcome the perfect prediction problem, Stata software includes the option "augment" (White et al. 2010). The "augment" option employs an augmented-regression. The is an ad hoc process suggested by White et al (2010). In the augment process, Stata creates a small number of additional observations that are added to the original data during estimation of model parameters. These extra observations assist in preventing perfect prediction. The extra observations are given a small weight to limit their impact on the estimates of important parameters such as effect size (White et al. 2010).

However, the simulations for this thesis actually indicated that, in practice, even the use of the option augment is not enough to completely prevent perfect prediction when both efficacy rates are close to boundary – but it does greatly improve the problem of perfect prediction. In all scenarios where perfect prediction has been reported in this thesis, the problem was considerably worse when the "augment" option was omitted than when the option was included.

6.7 Bias towards the null observed when wrong models are used (MAR and MCAR)

In situations where group is linked to missingness as well as outcome, it has been shown in the simulations in this thesis that omission of the group variable from the imputation process tends to bias the results towards the null. The reason is that if group is omitted from the imputation process, the imputed outcome values will not be associated with the

outcome while observed values will. The overall association between the group and outcome will be reduced producing bias that goes toward the null hypothesis. This will increase type II error thereby reducing the power of the study to detect the difference. A Complete case analysis maintains the association that exists between the outcome and the variable (group) that is linked to both outcome and missingness. Consequently the results from the complete case analyses are often better than those of MI when data are MAR or MCAR

6.8 Perfect prediction in MI procedures

It has been observed in simulation findings that when both efficacy rates are close to 100% (boundary value), the imputed models result in some of the models failing to produce output. This happens irrespective of whether the imputed values are binary or continuous. In fact, the perfect prediction may rise in any Generalized Linear Model (GL) that has a categorical outcome (White et al. 2010). In the case of the simulations foe this thesis, it is likely that the reason for the perfect prediction is that all the imputed values take the same values across the imputations. This results in zero between imputation variance and therefore calculation of the degrees of freedom is impossible as division by zero is impossible. The degrees of freedom v for calculation of confidence intervals and statistical tests from MI, is given as

$v = (p-1)\left[1+\frac{\bar{\vartheta}}{(1+p^{-1})\xi}\right]^2$, where ξ is the between imputation variance as given by Rubin (1976). So when ξ is zero, there is likely to be the perfect prediction problem.

White et al (2010) additionally suggest that the problem arises with standard errors that are extremely large, that reflect the approximately flat nature of the likelihood.

In order to overcome the perfect prediction problem, Stata software uses the option "augment" (White et al. 2010) to handle perfect prediction directly during imputation. The "augment" option employs an augmented-regression. The is an ad hoc process that was suggested by White et al (2010). In the augment process, stata creates some few additional observations that are added to the original data during estimation of model parameters. These extra observations assist in preventing perfect prediction. The extra observations are given a small weight to limit their impact on the estimates of the parameters (White et al. 2010).

Practically in the simulations for this thesis, it was observed that even the use of the option augment is not enough to completely prevent perfect prediction when both efficacy rates are close to boundary. It however greatly improves the problem of perfect prediction. In all scenarios where perfect prediction has been reported in this thesis, the problem was considerably worse when the "augment" option was omitted than when the option was included.

6.9 Practical implications

In practice, the mechanisms for handling missing outcome observations are complicated, particularly when the outcome is a binary variable and so, by definition, can take only a limited (in this case, just two) values. In some cases, there may be

enough evidence that data are either MAR, MCAR or MNAR, but in others, it may be absolutely difficult to tell the mechanism that may be creating the missing outcomes. Below is a summary of what may be done under the different practical situations.

What if binary outcome data is MAR or MCAR?

In practice, if one believes that the data are MAR or MCAR, the CC and multiple imputation will mathematically give the same results with minor differences in favour of CC, but, although the MI offers no statistical advantages, the intention to treat principle (ITT) which is the standard for analysis of RCTs probably over-ride this and MI should be analysis of first choice. However, a CC analysis must be performed as part of the "Per Protocol" analyses, firstly because MI has no place in "per protocol analysis" but secondly, and perhaps more importantly, because the CC yields unbiased estimates of effect.

What if binary outcome data is MNAR?

In a MNAR situation, neither a CC analysis nor any version of multiple imputation yields unbiased estimates of effect difference. To reduce this bias as much as possible, an MI approach is required with the imputation calculations based on all possible variables that are linked to missingness in the outcomes. The exception to this is if treatment group is itself linked to missingness in the outcome variable – in this situation, group membership should *not* be included in the imputation process as this will serve to increase bias in the effect size estimate. Covariates linked directly to the outcomes

provide useful information in imputing the missing outcomes, but variables that are associated with both the outcome and missingness tend to exaggerate bias if included in the imputation process. As previously stated, a detailed examination of the inclusion of different types of covariate was not within the limits of this thesis; further research is needed to examine in detail the complex inter-relationship between the relationship between covariates and missingness in the MNAR condition to fully confirm the above recommendations.

What if it is not known whether missing outcomes are MAR, MCAR or MNAR?

This is clearly the most difficult situation to resolve analytically, and there may not be a simple answer. There are, however, some simple rules that can be applied. Firstly, the statistician must obtain as much information as possible about the study design, and must examine the correlation structure between the covariates of interest and also between these covariates and the (available) outcomes to try to identify important mathematical relationships that might give some clues as to the missingness structure. Secondly, all relevant members of the trial team need to then meet to discuss the conduct of the study and to explore practical issues that arose during the preparation for and the conduct of the trial that might have affected missingness in the outcome variable.

In the likely event that there will still be uncertainty about the missingness mechanism even when the above actions have been exhausted, the statistician must perform a

sensitivity analysis, effectively repeating the primary analysis while varying the missingness assumptions in order to estimate the likely bias in the effect size estimate under different missingness assumptions. The trial team is then faced with the final challenge of comparing the different estimates obtained in the sensitivity analysis and intuitively selecting the "best" of these for publication. In reality, this may mean reporting a range of possible estimates, with an indication that this range covers likely positive and negative bias, with the true unbiased estimate lying somewhere within the range.

6.9.1 Suggestions for further research

Inevitably, several "further research questions" were generated by these simulation studies. Two of these are considered to be of particular importance.

What is causing the irregular bias patterns in the COPY method of the standard binomial risk difference regression model method with increasing number of copies?

This was a particularly striking and puzzling finding, for which there is no intuitively obvious answer. While it was not surprising to find that bias levels changed with increasing number of copies, and it was equally unsurprising that bias at first increased then decreased (i.e. that there was a turning point in the bias trend), it was totally unexpected that the bias trend had more than one turning point. The "plateau" trend found at large numbers of copies was predictable, but as part of a general upwards or

downwards exponential trend rather than following a second turning point in the bias trend plot.

Further research into this slightly anomalous finding will need to explore whether the rather irregular changes in bias levels with increasing copy numbers is a direct consequence of the binomial model formula, is it merely a consequence of the mathematics underlying the COPY method itself, is it a boundary problem (i.e. a consequence of one or both of the effect rates approaching 100%), or is there some other explanation.

Why is the power of the study the same for CC analyses and Multiple Imputation analyses when outcomes are MNAR, missingness is related to treatment group membership, and group is included in the imputation calculations?

The rather simple mathematical / algebraic explanation offered for the different bias trends in the CC, MI without group and MI with group analyses under the MNAR condition appear to predict this empirical finding from the relevant simulation exercises. This mathematical explanation dealt solely with the very simple case of missingness being related *only* to treatment group membership. Intriguingly, the simulations looked at more complex situations involving additional covariates and still found an apparent equality of biases for the MI analysis with group included in the imputations and the CC analyses. Was this just a coincidence? Would even more complex simulations break this apparent equality of bias? If the biases remain the same even under more complex

situations, why is this so? This is not an intuitively obvious finding – but if it is a true generalisable finding, it may reveal something important about the mechanism of the MI process in this very particular case of missingness being related to treatment group membership and this group membership information being used to inform the imputation process.

6.10 Summary conclusion and recommendations

The binomial regression model with an identity link function was found to be very susceptible to model failure when modeling risk (or efficacy) differences. Augmenting the standard binomial regression model with the COPY method was found to be an ineffective approach for achieving convergent and unbiased estimates of the risk differences, particularly if there was a requirement to adjust this estimate the influence of confounding covariates and/or factors.

Cheung's modified OLS with Huber-White robust standard errors was found to be a possible and attractive alternative method in situations where the standard binomial regression method does not converge. However, based on its superior statistical properties when samples are small, the binomial model with an identity link function is recommended as the statistical analysis method of first choice for estimating risk / efficacy differences, ***provided the model converges***. Otherwise the more reliable Cheung modified OLS method should be used. Both statistical methods should be

allowed for in the Statistical Analysis Plan sections of clinical trials where the estimation of risk / efficacy differences is considered important.

In the comparisons of the methods for dealing with missing (binary) outcomes, both the complete case (CC) analysis approach and the multiple imputation (MI) approach that included treatment group membership in the imputation calculations provided unbiased estimates of both unadjusted and covariate adjusted risk differences in situations where either the MAR or MCAR assumption was considered plausible. The statistical efficiency of the complete case analysis approach was found to be the same as, and often better than, that obtained using the multiple imputation approach (provided that group membership was included in the imputation calculations – if group membership is excluded from the imputation calculations, the resultant effect size estimate will definitely be biased). Complete case analysis would thus appear to be the method of choice for the primary analysis, particularly give the additional advantage that it is simple to apply and invariably default approach in most statistical software packages.

However, the complete case approach contravenes the intention-to-treat (ITT) principle recommended by, among others, the CONSORT Guidelines for the analysis and reporting of comparative clinical trials. In which case, an appropriate MI analysis is recommended as the analysis method of first choice when the missingness in a binary outcome can be considered to be MAR or MCAR, with a per protocol analyses based on the complete case approach. There is no place for multiple imputation in a per protocol analysis by definition. The two approaches used in tandem in this way then serve as a

sensitivity analysis tool to provide a detailed evaluation of the likely bias in the effect size estimate.

When missing binary outcome values have to be assumed to be MNAR, neither a complete case analysis nor a multiple imputation based analysis is valid for obtaining unbiased unadjusted or covariate adjusted risk / efficacy difference estimates. Furthermore, including treatment group membership in the imputation calculation when group was considered to be related to missingness was found to inflate rather than to moderate bias levels. However, under the MNAR condition the multiple imputation approach is recommended as the statistical analysis method of first choice provided all covariates and factors considered to be related to missingness (but not group membership) are included in the imputation process.

References

Allison, P. D. 2001. Missing data: quantitative applications in the social sciences. SAGE Publications, Thousand Oaks, London.

Altman, D. G. 2009. Missing outcomes in randomized trials: addressing the dilemma. Open. Med. 3:e51-e53.

Altman, D. G. and J. M. Bland. 2007. Missing data. BMJ. 334:424.

Altman, D. G., K. F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Gotzsche, and T. Lang. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann. Intern. Med. 134:663-694.

Andridge, R. R. and R. J. Little. 2010. A Review of Hot Deck Imputation for Survey Non-response. Int. Stat. Rev. 78:40-64.

Arinaitwe, E., T. G. Sandison, H. Wanzira, A. Kakuru, J. Homsy, J. Kalamya, M. R. Kamya, N. Vora, B. Greenhouse, P. J. Rosenthal, J. Tappero, and G. Dorsey. 2009. Artemether-lumefantrine versus dihydroartemisinin-piperaquine for falciparum malaria: a longitudinal, randomized trial in young Ugandan children. Clin. Infect. Dis. 49:1629-1637.

Axelson, O., M. Fredriksson, and K. Ekberg. 1994. Use of the prevalence ratio v the prevalence odds ratio as a measure of risk in cross sectional studies. Occup. Environ. Med. 51:574.

Bang, H. and J. M. Robins. 2005. Doubly robust estimation in missing data and causal inference models. Biometrics. 61:962-973.

Barros, A. J. and V. N. Hirakata. 2003. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC. Med. Res. Methodol. 3:21.

Bell, D. J., S. K. Nyirongo, M. Mukaka, E. E. Zijlstra, C. V. Plowe, M. E. Molyneux, S. A. Ward, and P. A. Winstanley. 2008. Sulfadoxine-pyrimethamine-based combinations for malaria: a randomised blinded trial to compare efficacy, safety and selection of resistance in Malawi. PLoS One. 3:e1578.

Blizzard, L. and D. W. Hosmer. 2006. Parameter estimation and goodness-of-fit in log binomial regression. Biom. J. 48:5-22.

Borrmann, S., T. Peto, R. W. Snow, W. Gutteridge, and N. J. White. 2008. Revisiting the design of phase III clinical trials of antimalarial drugs for uncomplicated Plasmodium falciparum malaria. PLoS Med. 5:e227.

Brasseur, P., P. Agnamey, O. Gaye, M. Vaillant, W. R. Taylor, and P. L. Olliaro. 2007. Efficacy and safety of artesunate plus amodiaquine in routine use for the treatment of uncomplicated malaria in Casamance, southern Senegal. Malar. J. 6:150.

Buck, S. F. 1960. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Royal Statistical Society. B22:302-306.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder. 2006. The design of simulation studies in medical statistics. Stat. Med. 25:4279-4292.

Carpenter, J. R. and M. G. Kenward. 2006. A comparison of multiple imputation and doubly robust estimation for analyses with missing data. Journal of Royal Statistical Society. 169:571-584.

Carpenter, J. R. and M. G. Kenward. 2007. Missing data in clinical trials - a practical guide. [Pamphlet] Available from: www.hta.nhs.uk/nihrmethodology/reports/1589.pdf

Carpenter, J. R., M. G. Kenward, and I. R. White. 2007. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat. Methods Med. Res. 16:259-275.

Case, L. D., G. Kimmick, E. D. Paskett, K. Lohman, and R. Tucker. 2002. Interpreting Measures of Treatment Effect
in Cancer Clinical Trials. The Oncologist. 7:181-187.

Chasela, C. S., M. G. Hudgens, D. J. Jamieson, D. Kayira, M. C. Hosseinipour, A. P. Kourtis, F. Martinson, G. Tegha, R. J. Knight, Y. I. Ahmed, D. D. Kamwendo, I. F. Hoffman, S. R. Ellington, Z. Kacheche, A. Soko, J. B. Wiener, S. A. Fiscus, P. Kazembe, I. A. Mofolo, M. Chigwenembe, D. S. Sichali, and C. M. van der Horst. 2010. Maternal or infant antiretroviral drugs to reduce HIV-1 transmission. N. Engl. J. Med. 362:2271-2281.

Cheung, Y. B. 2007. A modified least-squares regression approach to the estimation of risk difference. Am. J. Epidemiol. 166:1337-1344.

Collins, L. M., J. L. Schafer, and C. M. Kam. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol. Methods. 6:330-351.

Cox, D. R. 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological). 34:187-220.

Cox, D. R. and E. J. Snell. 1970. Analysis of binary data. 2nd ed. Chapman & Hall/CRC, London.

Crompton, P. D., B. Traore, K. Kayentao, S. Doumbo, A. Ongoiba, S. A. Diakite, M. A. Krause, D. Doumtabe, Y. Kone, G. Weiss, C. Y. Huang, S. Doumbia, A. Guindo, R. M. Fairhurst, L. H. Miller, S. K. Pierce, and O. K. Doumbo. 2008.

Sickle cell trait is associated with a delayed onset of malaria: implications for time-to-event analysis in clinical studies of malaria. J. Infect. Dis. 198:1265-1275.

Cummings, P. 2009a. The relative merits of risk ratios and odds ratios. Arch. Pediatr. Adolesc. Med. 163:438-445.

Cummings, P. 2009b. Methods for estimating adjusted risk ratios. The Stata Journal. 9:175-196.

Davies, H. T., I. K. Crombie, and M. Tavakoli. 1998. When can odds ratios mislead? BMJ 316:989-991.

Deddens, J. A. and M. R. Petersen. 2008. Approaches for estimating prevalence ratios. Occup. Environ. Med. 65:481, 501-481, 506.

Deddens, J. and M. R. Petersen. 2003. Estimation of prevalence ratios when PROC GENMOD does not converge, pp. 1-6. In: Proceedings of the 28th Annual SAS Users Group International Conference, 30 March-2 April 2003. Paper 270-28. Cary, NC: SAS Institute Inc, 2003. Available from: http://www2.sas.com/proceedings/sugi28/270-28.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. B39:1-38.

Dempster, A. P. and D. B. Rubin. 1983. Overview of incomplete data in sample surveys. In: Madow, W.G., I. Olkin, and D.B Rubin. Vol.II: Theory and annotated bibliography. New York, Academic Press.

Desai, M., D. A. Esserman, M. D. Gammon, and M. B. Terry. 2011. The use of complete-case and multiple imputation-based analyses in molecular epidemiology studies that assess interaction effects. Epidemiol. Perspect. Innov. 8:5.

Diggle, P. J., P. J. Heagerty, K. Liang, and S. L. Zeger. 2002. Analysis of longitudinal data. 2nd ed. Oxoford University Press, Oxford, New York.

Diggle, P. J. and M. G. Kenward. 1994. Informative dropout in longitudinal data analysis. Journal of the Royal Statistical Society Series C (Applied Statistics). 43:49-94. 1994.

Donders, A. R., G. J. van der Heijden, T. Stijnen, and K. G. Moons. 2006. Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. 59:1087-1091.

Enders, C. K. 2010. Applied missing data analysis. The Guilford Press, New York, London.

Faucett, C. L., N. Schenker, and J. M. Taylor. 2002. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. Biometrics. 58:37-47.

Faucher, J. F., A. Aubouy, A. Adeothy, G. Cottrell, J. Doritchamou, B. Gourmel, P. Houze, H. Kossou, H. Amedome, A. Massougbodji, M. Cot, and P. Deloron. 2009. Comparison of sulfadoxine-pyrimethamine, unsupervised artemether-lumefantrine, and unsupervised artesunate-amodiaquine fixed-dose formulation for uncomplicated plasmodium falciparum malaria in Benin: a randomized effectiveness noninferiority trial. J. Infect. Dis. 200:57-65.

French, N., S. B. Gordon, T. Mwalukomo, S. A. White, G. Mwafulirwa, H. Longwe, M. Mwaiponya, E. E. Zijlstra, M. E. Molyneux, and C. F. Gilks. 2010. A trial of a 7-valent pneumococcal conjugate vaccine in HIV-infected adults. N. Engl. J. Med. 362:812-822.

Gesase, S., R. D. Gosling, R. Hashim, R. Ord, I. Naidoo, R. Madebe, J. F. Mosha, A. Joho, V. Mandia, H. Mrema, E. Mapunda, Z. Savael, M. Lemnge, F. W. Mosha, B. Greenwood, C. Roper, and D. Chandramohan. 2009. High resistance of Plasmodium falciparum to sulphadoxine/pyrimethamine in northern Tanzania and the emergence of dhps resistance mutation at Codon 581. PLoS One. 4:e4569.

Graham, J. W. 2009. Missing data analysis: making it work in the real world. Annu. Rev. Psychol. 60:549-576.

Greenland, S. 1987. Interpretation and choice of effect measures in epidemiologic analyses. Am. J. Epidemiol. 125:761-768.

Grimes, D. A. and K. F. Schulz. 2008. Making sense of odds and odds ratios. Obstet. Gynecol. 111:423-426.

Groenwold, R. H., A. R. Donders, K. C. Roes, F. E. Harrell, Jr., and K. G. Moons. 2011. Dealing with missing outcome data in randomized trials and observational studies. Am. J. Epidemiol. 175:210-217.

Hedeker, D. and R. D. Gibbons. 2006. Longitudinal data analysis. Wiley-Interscience, New Jersey.

Higgins, J. P., I. R. White, and A. M. Wood. 2008. Imputation methods for missing outcome data in meta-analysis of clinical trials. Clin. Trials. 5:225-239.

Ibrahim, J. G. and G. Molenberghs. 2009. Missing data methods in longitudinal studies: a review. Test. (Madr.) 18:1-43.

Kenward, M. G. and J. Carpenter. 2007. Multiple imputation: current perspectives. Stat. Methods Med. Res. 16:199-218.

Klebanoff, M. A. and S. R. Cole. 2008. Use of multiple imputation in the epidemiologic literature. Am. J. Epidemiol. 168:355-357.

Lachin, J. M. 1999. Worst-rank score analysis with informatively missing observations in clinical trials. Control Clin. Trials. 20:408-422.

Lee, Y. J., J. H. Ellenberg, D. G. Hirtz, and K. B. Nelson. 1991. Analysis of clinical trials by treatment actually received: is it really an option? Stat. Med. 10:1595-1605.

Little, R. J. R. D. B. 2002. Statistical analysis with missing data. 2nd ed. Wiley, New York.

Liublinska, V. and D. B. Rubin. 2012. Re: "dealing with missing outcome data in randomized trials and observational studies". Am. J. Epidemiol. 176:357-358.

Lumley, T., R. Kronmal, and S. Ma. 2006. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper 293. Available from: http://www.bepress.com/uwbiostat/paper293

Machekano, R. N., G. Dorsey, and A. Hubbard. 2008. Efficacy studies of malaria treatments in Africa: efficient estimation with missing indicators of failure. Stat. Methods Med. Res. 17:191-206.

Magder, L. S. 2003. Simple approaches to assess the possible impact of missing outcome information on estimates of risk ratios, odds ratios, and risk differences. Control Clin. Trials. 24:411-421.

Marshall, A., D. G. Altman, R. L. Holder, and P. Royston. 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC. Med. Res. Methodol. 9:57.

Marshall, A., D. G. Altman, P. Royston, and R. L. Holder. 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC. Med. Res. Methodol. 10:7.

Mattei, A. 2009. Estimating and using propensity score in presence of missing background data: an application to assess the impact of child bearing on wellbeing. Statistical Methods and Applications. 18:257-273.

McNutt, L. A., C. Wu, X. Xue, and J. P. Hafner. 2003. Estimating the relative risk in cohort studies and clinical trials of common outcomes. Am. J. Epidemiol. 157:940-943.

Miettinen, O. S. and E. F. Cook. 1981. Confounding: essence and detection. Am. J. Epidemiol. 114:593-603.

Molenberghs, G. and M. Kenward. 2007. Missing data in clinical studies. John Wiley & Sons, Chichester, West Sussex.

Molenberghs, G., H. Thijs, I. Jansen, C. Beunckens, M. G. Kenward, C. Mallinckrodt, and R. J. Carroll. 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics. 5:445-464.

Mallinckrodt, and R. J. Carroll. 2004. Analyzing incomplete longitudinal clinical trial data. Biostatistics. 5:445-464.

Montori, V. M. and G. H. Guyatt. 2001. Intention-to-treat principle. CMAJ. 165:1339-1341.

Page, J. and J. Attia. 2003. Using Bayes' nomogram to help interpret odds ratios. ACP J. Club. 139:A11-A12.

Petersen, M. R. and J. A. Deddens. 2008. A comparison of two methods for estimating prevalence ratios. BMC Medical Research Methodology. 8.

Petersen, M. R. and J. A. Deddens. 2009. A revised SAS macro for maximum likelihood estimation of prevalence ratios using the COPY method. Occup. Environ. Med. 66:639.

Robins, J. M., A. Rotnitzky, and L. P. Zhao. 1995. Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association. 90:106-121.

Rosenbaum, P. T. and D. B. Rubin. 1983. The central role of propensity score in observational studies for causa effects. Biometrika. 70:41-55.

Rosenbaum, P. T. and D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensty score. Journal of the American Statistical Association. 79:516-524.

Rotnitzky, A. and J. Robins. 1997. Analysis of semi-parametric regression models with non-ignorable non-response. Stat. Med. 16:81-102.

Royston, P. and I. R. White. 2011. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. Journal of Statistical Software. 45:1-20.

Rubin D.B. 1976. Inference and missing data. Biometrika. 63:581-592.

Rubin D.B. 1996. Multiple imputation After 18+ Years. Journal of the American Statistical Association. 91:473-489.

Rubin, D. B. 1987. Multiple imputation for nonresponse in surveys. Wiley, New York.

SAS Technical Support. 2009. SAS/STAT(R) 9.2 user's guide. 2nd ed. Available from: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#whatsnew_toc.htm.

Savu, A., Q. Liu, and Y. Yasui. 2010. Estimation of relative risk and prevalence ratio. Stat. Med. 29:2269-2281.

Schafer, J. L. 1997. Analysis of incomplete multivariate data. Chapman & Hall/CRC Press, London. [Abstract]

Schafer, J. L. 1999. Multiple imputation: a primer. Stat. Methods Med. Res. 8:3-15.

Schafer, J. L. and J. W. Graham. 2002. Missing data: our view of the state of the art 43. Psychol. Methods. 7:147-177.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins. 1999. Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American StatisticalAssociation. 94:1096-1120.

Shapiro, S. 2001. The revised CONSORT statement: honing the cutting edge of the randomized controlled trial. CMAJ. 164:1157-1158.

Sterne, J. A., I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 338:b2393.

Streiner, D. L. 2008. Missing data and the trouble with LOCF. Evid. Based. Ment. Health 11:3-5.

van Buuren S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. 16:219-242.

van Buuren S., H. C. Boshuizen, and D. L. Knook. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. 18:681-694.

Wacholder, S. 1986. Binomial regression in GLIM: estimating risk ratios and risk differences. Am. J. Epidemiol. 123:174-184.

Walter, S. D. 2000. Choice of effect measure for epidemiological data. J. Clin. Epidemiol. 53:931-939.

White, I. R. and J. B. Carlin. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat. Med. 29:2920-2931.

White, I. R., R. Daniel, and P. Royston. 2010. Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical data. Computational Statistics and Data Analysis. 54:2267-2275.

Wood, A. M., I. R. White, and S. G. Thompson. 2004. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin. Trials. 1:368-376.

World Health Organization. 2003. Assessment and monitoring of antimalarial drug efficacy for the treatment of uncomplicated falciparum. World Health Organization, Geneva.

Zhang, J. and K. F. Yu. 1998. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA. 280:1690-1691.

Appendices

Stata programs

n.b. the programs were substituted by relevant parameters to achieve different scenarios

Appendix: A1 stata commands for generating MCAR data

Missing Completely at Random scenarios

```
cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data" // specifying directory

capture program drop RCT // to drop the program before running the updated

capture program RCT, rclass //

drop _all // to drop all matrix

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, ln(wt) and ln(para)

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200

gen age=exp(lnage) // transforming log age to original scale

gen wt=exp(lnwt) // transforming log  wt to original scale

gen para=exp(lnpara) // transforming log para to original scale


drop lnage lnwt lnpara // to drop variables that are on log scale

replace age=60 if age>60 // to maintain age eligibility criteria

replace age=12 if age<12 // to retain age eligibility criteria
```

```
gen studyno=_n  // studyno takes observation number

order studyno hb age wt para // ordering variables

gen rand=uniform()  // generate uniform distribution

gen block=int((studyno-1)/10)  // generate block as integer part

sort block rand // sorting block and rand

gen grp=1 if block!=block[_n-1] // to assign a block number

replace grp=grp[_n-1]+1 if block==block[_n-1] //assign block number

gen grpcode="A"  // assigning labels to groups

replace grpcode="B" if grp>=6  // assigning labels

gen group=. // generate blank variable called group

replace group=1 if grpcode=="A"  // replace value

replace group=0 if grpcode=="B" // replace value

gen result=.

gen result1=.

replace result1=rbinomial(1,0.85) // 0.85 is replaced by relevant probability value to achieve desired efficacy

gen result2=.

replace result2=rbinomial(1,0.60) // 0.60 is replaced by relevant probability value to achieve

replace result=result1 if group==0

replace result=result2 if group==1

tab result group, col

keep studyno hb age wt para group result

gen outcome=result

gen rand1=uniform() //random number generation

sort rand1  //to sort treatment in ascending then in each treatment to sort x in descending order

replace outcome=. in 1/10 //to create missing 5% MCAR
```

```
mi set mlong // setting multiple imputation procedure

mi register imputed outcome // register variable to be imputed

mi impute logit outcome wt hb age para, add(10) // impute 10 different data sets using
                MICE, logit is replaced by regress to to impute on continuous scale

mi estimate: regress outcome group hb age, vce(robust) //Use rubins rule to obtain
pooled estimates, Cheung's method with robust standard errors are used.

matrix a=e(b_mi)

matrix s=e(V_mi)

matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4]))

matrix D=a, z

svmat D, names(vvector)

return scalar LL=vvector1 - invnormal(0.975)*vvector5

return scalar UL=vvector1 + invnormal(0.975)*vvector5

sum outcome if group==0

return scalar m=r(mean)

sum outcome if group==1

return scalar z=r(mean)

sum outcome

return scalar q=r(mean)

end

set seed 23082

simulate  vvector1 vvector5 r(LL) r(UL) r(m) r(z) r(q), reps(5000): RCT

renvars  _sim_1- _sim_7 \group SE LL UL m z q

gen coverage=1 if  (-.25>= LL &  -.25<= UL)

replace coverage=0 if coverage==.

sum group SE LL UL coverage if q!=.

save wt_hb_age_para2V85_MCAR_bound5%log290313, replace
```

Appendix: A2 stata commands for generating MAR data

Missing at random

```
cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data" // specify directory

capture program drop RCT // clear program before running next one

capture program RCT, rclass

drop _all // clear matrix

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, ln(wt) and ln(para)

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200

gen age=exp(lnage) // transform to original form

gen wt=exp(lnwt) // transform to original form

gen para=exp(lnpara) // transform to original form

*drop lnage lnwt lnpara

replace age=60 if age>60  // to retain eligibility

replace age=12 if age<12 // retain eligibility

gen studyno=_n  // generate study number equal to observation number

order studyno hb age wt para // order variables

gen rand=uniform()  //generate uniform  distribution variable

gen block=int((studyno-1)/10) // generating block

sort block rand // sorting data

gen grp=1 if block!=block[_n-1] // generating block number
```

```
replace grp=grp[_n-1]+1 if block==block[_n-1]

gen grpcode="A"  // generating treatment label

replace grpcode="B" if grp>=6

gen group=.

replace group=1 if grpcode=="A"

replace group=0 if grpcode=="B"

gen result=.

gen result1=.

replace result1=rbinomial(1,0.85) // 0.85 is replaced by relevant probability value to achieve

gen result2=.

replace result2=rbinomial(1,0.60) // 0.60 is replaced by relevant probability value to achieve

replace result=result1 if group==0

replace result=result2 if group==1

tab result group, col

keep studyno hb age wt para group result

gen p2=invlogit(2*group+0.277*wt) //to generate a probability of missing as a function of weight and group resulting in MAR missingness

gen outcome= result

gen rand1=uniform() //random number generation

gen  x=p2*rand1 // this will assist in achieving a % of missing data of desired rate

gsort -x  //to sort treatment in ascending then in each treatment to sort x in descending order

replace outcome=. in 1/10 //to create missing 5% in group one with probability of missing depending on weight and group

mi set mlong // setting multiple imputation procedure

mi register imputed outcome // register variable to be imputed

mi impute logit outcome wt hb age para, add(10) // impute 10 different data sets using MICE, logit is replaced by regress to to impute on continuous scale
```

```
mi estimate: regress outcome group hb age, vce(robust) //Use rubins rule to obtain pooled estimates

matrix a=e(b_mi) // extract coefficient post estimation

matrix s=e(V_mi) // extract variance post estimation

matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4])) generate  matrix of standard errors

matrix D=a, z // generate matrix of coefficients and standard errors

svmat D, names(vvector) // generate vectors

return scalar LL=vvector1 - invnormal(0.975)*vvector5 // lower confidence limit

return scalar UL=vvector1 + invnormal(0.975)*vvector5 // upper confidence limit

sum outcome if group==0 // summarise data

return scalar m=r(mean) // extract mean from summary

sum outcome if group==1

return scalar z=r(mean)

sum outcome

return scalar q=r(mean)

end

set seed 23082 // setting a seed to reproduce data

simulate  vvector1 vvector5 r(LL) r(UL) r(m) r(z) r(q), reps(5000): RCT

renvars  _sim_1- _sim_7 \group SE LL UL m z q

gen coverage=1 if  (-.25>= LL &  -.25<= UL)

replace coverage=0 if coverage==.

sum group SE LL UL coverage if q!=.

save wt_hb_age_para2V85_bound5%log290313, replace
```

Appendix: A3 stata commands for generating MNAR data

Missing not at random

cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data" // specifying directory

capture program drop RCT // to drop the program before running the updated

capture program RCT, rclass //

drop _all // to drop all matrix

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, ln(wt) and ln(para)

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1, 0.09, 0.16, 0.02 \ 0.09, 1, 0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200

gen age=exp(lnage) // transforming log age to original scale

gen wt=exp(lnwt) // transforming log wt to original scale

gen para=exp(lnpara) // transforming log para to original scale

drop lnage lnwt lnpara // to drop variables that are on log scale

replace age=60 if age>60 // to maintain age eligibility criteria

replace age=12 if age<12 // to retain age eligibility criteria

gen studyno=_n // studyno takes observation number

order studyno hb age wt para // ordering variables

gen rand=uniform() // generate uniform distribution

gen block=int((studyno-1)/10) // generate block as integer part

sort block rand // sorting block and rand

```
gen grp=1 if block!=block[_n-1] // to assign a block number
replace grp=grp[_n-1]+1 if block==block[_n-1] //assign block number
gen grpcode="A"  // assigning labels to groups
replace grpcode="B" if grp>=6  // assigning labels
gen group=. // generate blank variable called group
replace group=1 if grpcode=="A"  // replace value
replace group=0 if grpcode=="B" // replace value
gen result=.
gen result1=.
replace result1=rbinomial(1,0.85) // generate Bernoulli  variable
gen result2=.
replace result2=rbinomial(1,0.60)
replace result=result1 if group==0
replace result=result2 if group==1
tab result group, col

keep studyno hb age wt para group result
gen outcome=result
gen miss2=rbinomial(1,0.07)
replace outcome=. if outcome==1 & miss2==1
egen count=count(miss2) if outcome==.

gen misspercent=count/_N
sort misspercent
return scalar x=misspercent in 1

mi set mlong // setting multiple imputation procedure
```

```
mi register imputed outcome // register variable to be imputed

mi impute logit outcome wt hb age para, add(10) // impute 10 different data sets using
                MICE,  logit is replaced by regress to impute outcome on continuous
                scale and the variables for the imputation models are replaced as
                required

mi estimate: regress outcome group hb age, vce(robust) //Use rubins rule to obtain
pooled estimates

matrix a=e(b_mi)

matrix s=e(V_mi)

matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4]))

matrix D=a, z

svmat D, names(vvector)

return scalar LL=vvector1 - invnormal(0.975)*vvector5

return scalar UL=vvector1 + invnormal(0.975)*vvector5

sum outcome if group==0

return scalar m=r(mean)

sum outcome if group==1

return scalar z=r(mean)

sum outcome

return scalar q=r(mean)

end

set seed 23082

simulate  vvector1 vvector5 r(LL) r(UL) r(m) r(z) r(q) r(x), reps(5000): RCT

renvars  _sim_1- _sim_8 \group SE LL UL m z q x

gen coverage=1 if  ( -.25> LL & -.25 < UL)

replace coverage=0 if coverage==.

sum group SE LL UL coverage if q!=.

save wt_hb_age_para_mnar5%, replace
```

```
cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data"

capture program drop RCT

capture program RCT, rclass

drop _all

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, wt and para

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200


gen age=exp(lnage)

gen wt=exp(lnwt)

gen para=exp(lnpara)

*drop lnage lnwt lnpara

replace age=60 if age>60

replace age=12 if age<12

gen studyno=_n

order studyno hb age wt para

gen rand=uniform()

gen block=int((studyno-1)/10)

sort block rand

gen grp=1 if block!=block[_n-1]

replace grp=grp[_n-1]+1 if block==block[_n-1]

gen grpcode="A"

replace grpcode="B" if grp>=6
```

```
gen group=.
replace group=1 if grpcode=="A"
replace group=0 if grpcode=="B"

gen result=.
gen result1=.
replace result1=rbinomial(1,0.85)
gen result2=.
replace result2=rbinomial(1,0.60)
replace result=result1 if group==0
replace result=result2 if group==1
tab result group, col

keep studyno hb age wt para group result
gen outcome=result
gen miss2=rbinomial(1,0.20)
replace outcome=. if outcome==1 & miss2==1
egen count=count(miss2) if outcome==.
gen misspercent=count/_N
sort misspercent
return scalar x=misspercent in 1

mi set mlong // setting multiple imputation procedure
mi register imputed outcome // register variable to be imputed
mi impute logit outcome wt hb age para, add(10) // impute 10 different data sets
mi estimate: regress outcome group hb age, vce(robust) //Use rubins rule to obtain pooled estimates
matrix a=e(b_mi) // extract coefficient post estimation
```

```
matrix s=e(V_mi) // extract variance post estimation

matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4])) generate  matrix of standard errors

matrix D=a, z // generate matrix of coefficients and standard errors

svmat D, names(vvector) // generate vectors

return scalar LL=vvector1 - invnormal(0.975)*vvector5 // lower confidence limit

return scalar UL=vvector1 + invnormal(0.975)*vvector5 // upper confidence limit

sum outcome if group==0 // summarise data

return scalar m=r(mean) // extract mean from summary

sum outcome if group==1

return scalar z=r(mean)

sum outcome

return scalar q=r(mean)


end

set seed 23082010

simulate  vvector1 vvector5 r(LL) r(UL) r(m) r(z) r(q) r(x), reps(5000): RCT

renvars  _sim_1- _sim_8 \group SE LL UL m z q x

gen coverage=1 if  ( -.25> LL & -.25 < UL)

replace coverage=0 if coverage==.

sum group SE LL UL coverage if q!=.

save wt_hb_age_para_mnar15%, replace


cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data"

capture program drop RCT

capture program RCT, rclass
```

```
drop _all

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, wt and para

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200


gen age=exp(lnage)

gen wt=exp(lnwt)

gen para=exp(lnpara)

*drop lnage lnwt lnpara

replace age=60 if age>60

replace age=12 if age<12

gen studyno=_n

order studyno hb age wt para

gen rand=uniform()

gen block=int((studyno-1)/10)

sort block rand

gen grp=1 if block!=block[_n-1]

replace grp=grp[_n-1]+1 if block==block[_n-1]

gen grpcode="A"

replace grpcode="B" if grp>=6

gen group=.

replace group=1 if grpcode=="A"

replace group=0 if grpcode=="B"
```

```
gen result=.
gen result1=.
replace result1=rbinomial(1,0.85)
gen result2=.
replace result2=rbinomial(1,0.60)
replace result=result1 if group==0
replace result=result2 if group==1
tab result group, col

keep studyno hb age wt para group result
gen outcome=result
gen miss2=rbinomial(1,0.40)
replace outcome=. if outcome==1 & miss2==1
egen count=count(miss2) if outcome==.

gen misspercent=count/_N
sort misspercent
return scalar x=misspercent in 1

mi set mlong // setting multiple imputation procedure
mi register imputed outcome // register variable to be imputed
mi impute logit outcome wt hb age para, add(10) // impute 10 different data sets
mi estimate: regress outcome group hb age, vce(robust) //Use rubins rule to obtain pooled estimates
matrix a=e(b_mi) // extract coefficient post estimation
matrix s=e(V_mi) // extract variance post estimation
matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4])) generate  matrix of standard errors
```

```
matrix D=a, z // generate matrix of coefficients and standard errors
svmat D, names(vvector) // generate vectors
return scalar LL=vvector1 - invnormal(0.975)*vvector5 // lower confidence limit
return scalar UL=vvector1 + invnormal(0.975)*vvector5 // upper confidence limit
sum outcome if group==0 // summarise data
return scalar m=r(mean) // extract mean from summary
sum outcome if group==1
return scalar z=r(mean)
sum outcome
return scalar q=r(mean)
end
set seed 23082010
simulate  vvector1 vvector5 r(LL) r(UL) r(m) r(z) r(q) r(x), reps(5000): RCT
renvars  _sim_1- _sim_8 \group SE LL UL m z q x
gen coverage=1 if  ( -.25> LL & -.25 < UL)
replace coverage=0 if coverage==.
sum group SE LL UL coverage if q!=.
save wt_hb_age_para_mnar30%, replace
```

Appendix: A4 stata commands for the Copy method and convergence

COPY method and convergence

```
cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data"

set more off

capture program drop RCT

capture program RCT, rclass

drop _all

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, ln(wt) and ln(para)

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200

gen age=exp(lnage)

gen wt=exp(lnwt)

gen para=exp(lnpara)

*drop lnage lnwt lnpara

replace age=60 if age>60

replace age=12 if age<12

gen studyno=_n

order studyno hb age wt para

gen rand=uniform()

gen block=int((studyno-1)/10)

sort block rand
```

```
gen grp=1 if block!=block[_n-1]
replace grp=grp[_n-1]+1 if block==block[_n-1]
gen grpcode="A"
replace grpcode="B" if grp>=6
gen group=.
replace group=1 if grpcode=="A"
replace group=0 if grpcode=="B"
gen result=.
gen result1=.
replace result1=rbinomial(1,0.80)
gen result2=.
replace result2=rbinomial(1,0.60)
replace result=result1 if group==0
replace result=result2 if group==1
tab result group, col
keep studyno hb age wt para group result
binreg result group age, rd iterate(1600) // replace with other variables as necessary
matrix a=e(b)
matrix s=e(V)
matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4]))
matrix D=a, z
svmat D, names(vvector)
return scalar LL=vvector1 - invnormal(0.975)*vvector5
return scalar UL=vvector1 + invnormal(0.975)*vvector5
end
set seed 23082
simulate  vvector1 vvector5 r(LL) r(UL), reps(5000): RCT
```

```
renvars  _sim_1- _sim_4 \group SE LL UL

gen coverage=1 if  (-.20>= LL &  -.20<= UL)

replace coverage=0 if coverage==.

sum group if group!=.

save age_converge, replace
```

n.b. the programs were substituted by relevant parameters to achieve different scenarios

Appendix: A5 stata commands for Cheung's OLS method and convergence

Cheung's OLS method and convergence

```
cd "C:\Documents and Settings\mmukaka\My Documents\Backup\PhD Files\Missing data"

set more off // to suppress more in stata

capture program drop RCT

capture program RCT, rclass

drop _all

matrix m = (3.15, 9.32, 2.4,10.7) // means for ln(age), hb, ln(wt) and ln(para)

matrix sd =(0.42,1.66, 0.18, 1.5) // sds for ln(age), hb and ln(wt) and ln(para) respectively

matrix C=(1,  0.09, 0.16, 0.02 \ 0.09, 1,  0.4, 0.2 \ 0.16, 0.4, 1, 0.05 \0.02, 0.2, 0.05, 1) //this is the correlation matrix for ln(age), hb, ln(wt), ln(para)

matrix list C //to check if the correlation matrix has been set up properly

drawnorm lnage hb lnwt lnpara, n(200) means(m) sds(sd) corr(C) //generating a multivariate normal distribution of sample size of 200

gen age=exp(lnage)

gen wt=exp(lnwt)

gen para=exp(lnpara)

*drop lnage lnwt lnpara

replace age=60 if age>60

replace age=12 if age<12

gen studyno=_n

order studyno hb age wt para

gen rand=uniform()

gen block=int((studyno-1)/10)

sort block rand

gen grp=1 if block!=block[_n-1]
```

```
replace grp=grp[_n-1]+1 if block==block[_n-1]
gen grpcode="A"
replace grpcode="B" if grp>=6
gen group=.
replace group=1 if grpcode=="A"
replace group=0 if grpcode=="B"
gen result=.
gen result1=.
replace result1=rbinomial(1,0.80)
gen result2=.
replace result2=rbinomial(1,0.60)
replace result=result1 if group==0
replace result=result2 if group==1
tab result group, col
keep studyno hb age wt para group result
regress result group age, rd iterate(1600) // replace with other variables as necessary
matrix a=e(b)
matrix s=e(V)
matrix z=(sqrt(s[1,1]), sqrt(s[2,2]), sqrt(s[3,3]), sqrt(s[4,4]))
matrix D=a, z
svmat D, names(vvector)
return scalar LL=vvector1 - invnormal(0.975)*vvector5
return scalar UL=vvector1 + invnormal(0.975)*vvector5
end
set seed 23082
simulate  vvector1 vvector5 r(LL) r(UL), reps(5000): RCT
renvars  _sim_1- _sim_4 \group SE LL UL
```

gen coverage=1 if (-.20>= LL & -.20<= UL)

replace coverage=0 if coverage==.

sum group if group!=.

save age_converge, replace

n.b. the programs were substituted by relevant parameters to achieve different scenarios