University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2005 *Paper* 193

Efficacy Studies of Malaria Treatments in Africa: Efficient Estimation with Missing Indicators of Failure

Rhoderick N. Machekano* Grant Dorsey†

Alan E. Hubbard‡

*Division of Biostatistics, School of Public Health, University of California, Berkeley, rodmach@berkeley.edu

†School of Medicine, University of California, San Francisco, grantd@itsa.ucsf.edu

‡Division of Biostatistics, School of Public Health, University of California, Berkeley, hubbard@stat.berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper193

Efficacy Studies of Malaria Treatments in Africa: Efficient Estimation with Missing Indicators of Failure

Rhoderick N. Machekano, Grant Dorsey, and Alan E. Hubbard

Abstract

Efficacy studies of malaria treatments can be plagued by indeterminate outcomes for some patients. The study motivating this paper defines the outcome of interest (treatment failure) as recrudescence and for some subjects, it is unclear whether a recurrence of malaria is due to that or new infection. This results in a specific kind of missing data. The effect of missing data in causal inference problems is widely recognized. Methods that adjust for possible bias from missing data include a variety of imputation procedures (extreme case analysis, hot-deck, single and multiple imputation), inverse weighting methods, and likelihood based methods (data augmentation, EM procedures and their extensions). In this article, we focus on multiple imputation, two inverse weighting procedures (the inverse probability of censoring weighted (IPCW) and the doubly robust (DR) estimators), and a likelihood based methodology (G-computation), comparing the methods' applicability to the efficient estimation of malaria treatments effects. We present results from a simulation study as well as results from a data analysis of malaria efficacy studies from Uganda.

1 Introduction

Randomized controlled trials are pivotal to public health and medical decisions because they have the potential to produce unbiased estimates of causal effects, allowing appropriate evaluation of competing treatments or interventions. However, many trials often suffer from a number of complications, among them, missing outcomes. Such data pathologies affect the estimation of causal effects, leading to biased estimates if not properly addressed. A recent review of published randomized controlled trials in major medical journals established that missing outcome data are a common problem and are often inadequately handled in statistical analyses [16].

The statistical framework for causal inference that has been widely adopted is one based on potential outcomes, originally introduced by Neyman (1923) for randomized experiments, and generalized and extended by Rubin to non-randomized studies. Now known as the "Rubin Causal Model" [5], a study unit (e.g., a patient), has the potential to be given any of the experimental treatments, and associated with the treatments are the potential outcomes, defined as all the outcomes that would be observed when each of the treatments would be applied to each of the units [1]. The causal effect between two treatments is then a comparison of the potential outcomes of the same group of units under the two treatment conditions.

Despite extensive studies of causal inference in the presence of missing data , statistical practice often excludes observations with any missing outcomes, simplifying the analysis at the expense of increased bias and reduced efficiency due to reduced sample sizes. Approaches that exclude observations with missing values (known as *complete case analysis*) ignore the possibility of existence of systematic differences between complete data cases and incomplete data cases, and that the resulting inference may not be generalizable to the population, especially when the degree of incompleteness is high. Another approach to analysis of missing data is to impute missing values with worst-possible (or best-possible) values (also known as *extreme case analysis*). This approach has no scientific merit except as a form of sensitivity analysis [5]. Other imputation methods such as Rubin's multiple imputation [4] are more attractive under certain assumptions on the missing data mechanism.

Valid inferences in the presence of missing outcome data, require methods that remove the bias introduced by the association of missingness mechanism to outcomes. Techniques that make assumptions about how the probability of an outcome being missing relates to covariates and outcomes have been proposed, and these can yield unbiased and efficient estimates under varying missingness types.

Malaria drug efficacy studies provide typical examples where missing outcomes are common. In these studies, the objective is to compare the efficacy of different

antimalarial treatments in curing malaria infection. The outcome, defined as the success or failure of antimalarial drugs in eliminating malaria parasites in infected individuals, can be clinically undetermined. In highly endemic areas where recurrent disease following malaria therapy is common [13], it is clinically impossible to distinguish recrudescence (true treatment failure) from new infections. Molecular genotyping techniques have been used to distinguish between re-infection and recrudescence [14]. The principle behind molecular genotyping is to compare genotyping patterns based on highly polymorphic genes in pre- and post-treatment samples in patients with recurrent malaria following treatment. Post-treatment samples containing only parasite strains present in pre-treatment samples are generally classified as recrudescence of resistant parasites. Post-treatment samples containing only new parasite strains not present in pre-treatment samples are generally classified as newly acquired infections. However, a complex situation arises when post-treatment samples contain some strains present in pre-treatment samples as well as new strains not present in the pre-treatment samples - a mixed genotype result. Classifying mixed genotyping results as either due to recrudescence or new infections likely over- or under-estimates the true risk of treatment failure respectively. In circumstances like these the mixed genotype is regarded as missing outcomes, complicating estimation of treatment effects.

In this article we explore methods suitable for the estimation of treatment effects in malaria studies in the presence of indeterminate outcomes (i.e. mixed genotypes). We start by describing the structure of the data and notation we use, and reviewing the types of missingness plausible in malaria studies. In section 2 we present methods that can be used to estimate our parameter of interest, treatment specific failure rate. In particular, we review two inverse probability weighting methods that have been developed as tools to remove potential bias introduced through unequal distribution of some factors (e.g. treatment, censoring and missingness). We also review the multiple imputation procedure, a widely used likelihood-based technique for analysis of missing data due to Rubin et al [4]. Finally, we will present the G-computation method adapted to missing data problems. In section 3, we compare estimates from the three procedures in addition to the extreme case (EC) analyses, in which all missing data are given either the worst case outcome (i.e. mixed genotype = failure), or the best case outcome (mixed genotype = non-failure) and the complete case (CC) analyses using simulated data. We present results comparing the six approaches to a full data (FD) analysis in which missingness in absent. In section 4, we apply the methods to malaria-treatment efficacy studies from Uganda. We conclude the paper with a discussion of our observations from the simulations and application to the data. The R code used to run the simulations is included in the appendix.

1.1 Data structure and notation

Let Y be the outcome of interest (in malaria studies,$Y = 1$ if recrudescence otherwise $Y = 0$), W the vector of measured covariates and A the k-level treatment variable. Let δ be the missingness indicator taking the value 1 if the outcome Y is observed and 0 otherwise. The data we discuss consist of n iid copies of the observation $O = (W, A, \delta, \delta Y)$. Let $X = (Y_a, W), \forall a \in \mathcal{A}$, where $\mathcal{A}$ is the support of A, define the full data structure, where Y_a is the counterfactual outcome for treatment $A = a$.

Our goal is to use the observed data $O = \{O_i : i = 1, \cdots, n\}$ to estimate the treatment specific failure rate $p_a = E(Y_a)$ or functions of p_a (e.g. Risk Difference $p_{a_1} - p_{a_0}$ or Risk Ratio $\frac{p_{a_1}}{p_{a_0}}$). Estimation methods of p_a are associated with certain assumptions on the underlying missingness pattern. We briefly state below the different types of missingness.

1.2 Missingness Mechanisms

Rubin [9] defined the following missingness patterns based on how missing variable indicators are related to the underlying values of the variables in the dataset. We can describe the missing data mechanism by the conditional distribution of missingness indictor, δ, given the complete data,$X^{comp} = (W, A, Y)$, where data from every subject in the study is completely observed. Define $\pi(W, A, Y) = P(\delta = 1|W, A, Y)$ and $\overline{\pi(W, A, Y)} = P(\delta = 0|W, A, Y)$. Following Rotnitzky, Robins and Scharfstein [8], we denote the model for the conditional probability of δ given A, W and Y by:

$$logit(\overline{\pi(W, A, Y)}) = \eta(A, W) + q(Y), \tag{1}$$

where $\eta(A, W)$ is an unknown frunction of A and W, and $q(Y)$ is a specified function of Y.

Data are said to be "missing completely at random" (MCAR) if the probability of a missing outcome does not dependent on measured covariates and outcomes i.e. $\overline{\pi(W, A, Y)} = c$ for some constant $c \in (0, 1)$. When $q(Y)$ equals 0, the data are said to be "missing at random" (MAR) i.e. the probability of missingness depends only on measured covariates A and W. Data MAR are also said to be "coarsened at random" (CAR) [3], [2]. Finally, data are "missing not at random" (MNAR) when missingness additionally depends on unobserved data Y, $(q(Y) \neq c, c \in \mathcal{R})$ after conditioning on the observed data.

2 Estimation of p_a

In this section, we discuss estimation of treatment specific failure rate, p_a, when treatment is randomized. Treatment assignment is said to be completely randomized if, given the full data X, $E[Y|A = a] = E[Y_a]$. The general case covering non-randomized studies is presented in the appendix. In the following discussions we assume model 1 with $q(Y) = 0$ i.e. data are MAR or MCAR.

Although we are interested in estimating $E[Y_a]$, the covariates W can be used in the estimation procedure to increase efficiency and reduce bias.

Let

$$E[Y|A = a] = m(a, \beta) \tag{2}$$

where m is some known non-negative function and β is a $k \times 1$ vector of unknown parameters. In the absence of missing data, estimation of $E(Y_a)$ can easily be achieved by ordinary regression procedures. When the outcome Y is binary, the logit function is a reasonable choice for m.

When data are assumed to be MCAR, the complete case analysis using ordinary regression procedures (regressing Y on A) gives unbiased estimates of $E(Y_a)$ [15]. However, if the degree of missingness is high, ordinary regression on complete cases can result in inefficient estimates. When data are MAR, estimates from ordinary regression of Y on A can be biased and inefficient [15], [4]. The extent of the bias depends on the proportion missing and the strength of the relationship between missingness and covariates that maybe related to the outcome. Robust methods are needed.

2.1 General Inverse Weighted Estimators

The class of inverse-weighted estimators have been extensively studied by Robins and colleagues [15], [6], [7], [12]. The inverse weighted estimators are motivated by ideas from efficiency theory of semiparametric models. Instead of using likelihood based estimation to estimate parameters of interest, a general estimating function methodology is preferred because it takes into account the coarsening mechanism (known or estimated) and has been shown to produce estimates that perform better in high-dimensional data sets with realistic sample sizes. These estimators seem to tie in with ideas from survey theory that weighting observations by their probability of missingness creates a pseudo-population in which missingness is no longer associated with failure.

Consider the following class of estimating functions:

$$D(O,h,\beta,\phi) = \frac{\delta h(A)\epsilon(\beta)}{\pi(W,A,Y)} + \left\{1 - \frac{\delta}{\pi(W,A,Y)}\right\}\phi(W,A,Y) \tag{3}$$

where $\epsilon(\beta) = Y - m(A|\beta)$ and h is some function of A and ϕ is some function of the observed data. We have the following properties:

1. Regardless of the choice of ϕ, $D(O,h,\beta,\phi)$ has mean 0 provided missingness model is correctly specified.

2. If $\phi(W,A,Y) = h(A)(Q(A,W) - m(a,\beta))$, where $Q(A,W)$ is some estimate of $E[Y|A,W]$, then $D(O,h,\beta,\phi)$ has mean 0 if either the missingness model or $Q(A,W)$ is correctly specified.

2.2 Inverse Probability of Censoring Weighted Estimator (IPCW)

In IPCW estimation, each complete observation is weighted with the inverse of the probability of being observed and used in solving a set of estimating equations. Consistent estimation of p_a or β requires the estimation of the missingness mechanism, $\pi(W,A,Y)$.

Assuming missingness is conditionally independent of failure Y given treatment A and covariates W (CAR), $\delta \perp Y|(A,W)$, implies missingness is determined only by the assigned treatment and measured covariates and not on the unobserved outcome variable Y. We further assume that $\pi(A,W) > 0$ for all A and W.

From (3) above, let $\phi(W,A,Y) = 0$. The resulting estimating functions

$$\left\{D_h(O|\beta,g) = \frac{\delta h(A)\epsilon(\beta)}{\pi(W,A,Y)} : h\right\} \tag{4}$$

satisfy properties 1 and 2 above. If the MAR assumption holds, and there is positive probability of observing an outcome Y given covariates A,W, then for all $h \in \mathcal{H} = \{h(A)\}$ we have the following:

$$E\left(\frac{\delta h(A)\epsilon(\beta)}{\pi(W,A,Y)}\right) = 0, \tag{5}$$

Proof: By the MAR assumption,

$$\begin{aligned} E\left(\frac{\delta h(A)\epsilon(\beta)}{\pi(W,A,Y)}\right) &= E\left(\frac{\delta h(A)\epsilon(\beta)}{\pi(X)}\right) \\ &= E\left(\frac{\delta h(A)(Y-m(A,\beta))}{\pi(X)}\right) \\ &= E\left(E\left(\frac{\delta h(A)(Y-m(A,\beta))}{\pi(X)}|X\right)\right) \\ &= E\left(\frac{h(A)(Y-m(A,\beta))}{\pi(X)}|X\right)E(\delta|X) \\ &= E\left(h(A)(Y-m(A,\beta))|A\right) \\ &= 0 \end{aligned}$$

Thus, from equation (5), we have unbiased estimating equations of the form,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(A_i)\epsilon_i(\beta)}{\pi(A_i,W_i)} = 0, \tag{6}$$

Solving equation (6) gives the inverse probability of censoring weighted (IPCW) estimator, $\hat{\beta}_{IPCW}$ and is a consistent estimator of the true β for all $h \in \mathcal{H}$. One choice for h is $h^* = \frac{d}{d\beta}m(A,\beta)Var(\epsilon)^{-1}$. The implementation of the IPCW estimator proceeds by first estimating the missingness mechanism $pi(A,W)$ and then using the inverse estimated probabilities of missingness as weights in an ordinary regression. Since missingness is a binary variable, the probability of a missing outcome is estimated using logistic regression. For consistency of the IPCW estimator, the missingness mechanism π has to be correctly specified. We can gain efficiency from using a bigger model for $\pi(A,W)$, for example, by using general additive models or other modeling procedures [11]. Even if the Y is MCAR, we can model missingness as a function of covariates W to gain efficiency.

2.2.1 Inference for IPCW estimator

If the MAR assumption holds and the missingness mechanism is correctly specified, and if the distribution of the outcome Y is correctly specified by $m(A,W,\beta)$, then β_{IPCW} is consistently asymptotically normal (CAN) i.e.

$$\hat{\beta}_{IPCW} - \beta \approx \frac{1}{n}\sum_{i=1}^{n} IC(W_i,\delta_i,Y_i) \tag{7}$$

where IC is the influence curve.

Under regularity conditions, $\sqrt{n}(\hat{\beta}_{IPCW} - \beta)$ converges in distribution to a standard multivariate Normal Distribution with variance $E[IC(O,\delta)^T IC(O,\delta)]$ [15]. We can obtain 95% confidence intervals for $\hat{\beta}_{IPCW}$ using the .95-quantiles from the standard normal distribution and estimated standard errors from the diagonal elements of $E[\widehat{IC(O,\delta)}^T \widehat{IC(O,\delta)}]$.
If π_0 (true) is known, the the influence curve is

$$IC_h(O) = -\frac{d}{d\beta} E(D_h(O,\beta|\pi_0))^{-1}|_{\beta=\beta_0} D_h(O,\beta_0|\pi_0) \tag{8}$$

This gives conservative inference if π is estimated with an efficient estimator. Otherwise standard error estimates for $\hat{\beta}_{IPCW}$ can be obtained using bootstrap methods.

2.3 IPCW Doubly Robust (IPCW-DR) Estimator

The consistency of the IPCW estimator depends on the correct specification of the missingness mechanism,$\pi(W,A,Y)$, otherwise the parameter estimates can be inconsistent. The IPCW estimator does not make use of all available data since subjects with missing outcomes do not contribute. The efficiency of the IPCW estimator can be improved by subtracting its projection on a nuisance score tangent space [15].

Again, equation (3) and properties 1 and 2 imply

$$\sum_{i=1}^{n} D_h(O_i|\hat{\beta}_{DR},\hat{\pi},\hat{\phi}) = 0 \tag{9}$$

for each $h \in \mathcal{H}$ and π. By setting $\phi(W,A,Y) = h(A)(Q(A,W) - m(a,\beta))$, where $Q(A,W)$ is a model for$E[Y|A,W]$, equation (9) holds if *either* the missingness model or $Q(A,W)$ is correctly specified. Given MAR assumption, $E[Y|A,W] = E[Y|A,W,\delta=1]$ and thus we can estimate $Q(A,W)$ using a regression of Y on A and W only on the observed data ($\delta = 1$). Solving the above equation gives the doubly robust estimator $\hat{\beta}_{DR}$. The double robustness of the above estimate comes from the fact that if $Q(A,W)$ or missingness mechanism π is correctly specified, the estimate $\hat{\beta}_{DR}$ is consistent and asymptotically normal (CAN) [15]. Moreover, if both the failure model and the missingness mechanism are correctly specified $\hat{\beta}_{DR}$ is asymptotically efficient if the model for $m(a|\beta)$ is nonparametric.

If the failure model and the missingness model are correctly specified, the influence function of $\hat{\beta}_{DR}$ is $[\frac{d}{d\beta}E(D(O,h,\alpha,\beta,\phi))]^{-1}D(O,h,\alpha,\phi)$ and the asymptotic

variance is given by $E[D(O|\beta,\pi,\phi)\hat{}^T D(O|\beta,\pi,\phi)]$. However, when either model is misspecified, the variance of $\hat{\beta}_{DR}$ is most easily estimated through bootstrapping.

2.4 Multiple Imputation (MI) Procedure

The multiple imputation procedure was introduced by Rubin [10], [4]. Multiple imputation is a simulation based approach, where each missing value Y_{mis} is replaced with a vector of $D > 1$ plausible values, accounting for the uncertainty about the right value to impute. The multiply imputed data sets are analyzed using standard procedures for complete data and combining the results from these analyses.

Our problem of estimating treatment effects in the presence of extra covariate information has been called "uncongeniality" of the imputer's and the user's models i.e. the imputation and analysis models are different. Our imputation model involves fitting a logistic regression to estimate $E[Y = 1|A, W]$ using only the observed data. From the resulting model, we estimate the probability of failure given covariates $\hat{P}[Y = 1|A, W]$ for all the observations. We then impute the outcome Y for the missing cases from these estimated probabilities. Using the completed data we estimate that parameter of interest $p_a = E[Y_a]$ by fitting the usual logistic regression of Y on A. After sufficient repeated impute-estimation steps, we get an estimate $\hat{\beta}_{MI}$ by summarizing the D $\hat{\beta}$s i.e.

$$p_a^{MI} = \frac{1}{D}\sum_{d=1}^{D} p_a^{(d)}$$

If one does an infinite number of imputations, the multiple imputation procedure converges to:

$$\hat{E}(Y_a) = \frac{\sum_{i=1}^{n} I(A_i = a) * \left[\Delta_i * Y_i + (1 - \Delta_i) * \hat{E}(Y_i \mid A = a, W_i)\right]}{\sum_{i=1}^{n} I(A_i = a)} \qquad (10)$$

where $\hat{E}(Y_i \mid A = a, W_i)$ is based on the model used to do the imputations.

2.5 G-computation

Under the randomization assumption (RA) and missing at random (MAR) assumption, the likelihood of the observed data can be factored into a part that depends only on the distribution of the full data X, and a part that equals the missingness

mechanism π. Our parameter of interest can be calculated from the G-computation formula as follows:

$$\begin{aligned} E(Y_a) &\overset{\text{RA}}{=} E_W(E(Y_a|A=a,W)) \\ &\overset{\text{MAR}}{=} E_W(E(Y|A=a,W)) \\ &\overset{\text{estimated}}{\simeq} \frac{1}{n}\sum_{i=1}^{n}\hat{E}(Y|A=a,W_i) \end{aligned}$$

Thus the maximum likelihood estimate (MLE) is simply $E_W[W[Y|A,W,\delta = 1]]$. This procedure contrasts with the MI procedure since they both rely on the consistency of the estimate $\hat{E}[Y|A,W]$ and G-computation is more efficient. One can think of the MI procedure as part G-computation (over the missing data) and part simple averages of the Y's, but only over the data where $A = a$. If one assumes that the model $Q(A,W)$ is consistent, the G-computation estimate must be more efficient, even after an infinite number of imputations. However, if $Q(A,W)$ is misspecified the MI approach might be less biased if the proportion of missing data is small.

3 Simulation Studies

In this section, we carry out simulation studies to compare estimators of the parameter p_a and β obtained by:

- conventional regression analysis on complete cases only (CC).
- conventional regression on all cases where missing values have been imputed using multiple imputation procedures.
- G-computation in which the parameter of interest is estimated by averaging over the covariate.
- the IPCW estimator discussed in section (2.1) in which the naive estimating equation is weighted according to estimated probability of missingness given the observed data.
- the doubly robust estimator obtained by solving missingness-orthogonalized estimating functions
- conventional regression analysis assuming all missing cases are non-failures.

- conventional regression analysis assuming all missing cases are failures.

The simulations are motivated by malaria drug efficacy studies in which we would like to estimate and compare treatment specific risks of failure when some outcomes are indeterminate. We study the bias and efficiency of the estimators under different missingness mechanisms and fitted models.

For each of the n observations, we generate a normally distributed covariate $W \sim N(0,1)$, a Bernoulli treatment variable A randomized with probability $Pr(A = 1|W) = 0.5$, and a binary outcome variable Y with the logit of the probability of failure equal to

$$logit[P(Y = 1|A = a, W = w)] = \beta_0 + \beta_1 a + \beta_2 w + \beta_3 a \times w \tag{11}$$

at selected values of $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$. Assuming outcomes are missing at random, we generate a missingness indicator δ for each observation from the following Bernoulli distribution:

$$logit[P(\delta = 1|\alpha, W, A)] = \alpha_0 + \alpha_1 A + \alpha_2 W + \alpha_3 A \times W \tag{12}$$

The outcome Y is missing when $\delta = 0$ and Y is observed when $\delta = 1$.

Setting $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (-2, 2, 2, 0)$ 37% of the cases have missing outcomes - 30% missingness in treatment arm $A = 0$ and 45% missingness in treatment arm $A = 1$. Choosing $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, -2, 2, -1)$, we have a a 50% (p_0) failure rate in the treatment arm $A = 0$ and a 16% (p_1) failure rate in the treatment arm $A = 1$.

In the next set of simulations, we test the robustness of the estimators to misspecification of the conditional distribution of Y given treatment and covariates and misspecification of the missingness mechanism. We misspecify the missingness model in estimating weights by modeling the probability of being observed as a function of treatment only.

For each set of simulations, we generated 1000 data sets of sample size $n = 200$. The seven estimators of treatment effect were compared with the complete data (without missingness) estimate by looking at the bias, standardized bias and the root mean squared error. The standardized bias (STDBIAS) is given by

$$STDBIAS = 100 \times \frac{bias}{standarderror}$$

where bias is the deviation of the average estimate over the 1000 simulated data sets from the true parameter and the standard deviation is the average deviation of the 1000 estimates from their mean over the 1000 simulated data sets. The simulation results are summarized in tables (1), (2), and (3) from which we conclude the following:

1. Under the correct conditional distribution of Y given treatment A and covariates W, and the missingness mechanism $\pi(A, W)$, the inverse weighted methods (IPCW, DR), G-computation and multiple imputation have less bias and lower variance compared to the complete case and extreme case estimators.

2. Misspecifying conditional distribution of $Y|A, W$ extremely biases G-computation and multiple imputation, but not the Doubly Robust estimate

3. Misspecifying the missingness mechanism $\pi(X^{comp})$ biases the IPCW estimate more than it does the DR estimate

4. The risk differences do not differ substantially between estimators except under the extreme case failures assumption.

5. Assuming all missing outcomes are failures, the causal effect is very biased, indicating that there is no difference between treatments.

6. Overally, the Doubly Robust estimator is best since it is robust to model misspecifications

7. The G-computation performs similarly to the Multiple Imputation procedures, and since it is more efficient and requires the same assumptions (i.e. correct model for $E[Y|A, W]$), it is our preferred method.

4 Analysis of Uganda Malaria Data

We applied the methods discussed above to malaria data from randomized studies in Uganda. The studies were designed to compare the effectiveness of three antimalarial drug regimens. Malaria infected patients from regions of different transmission intensities were randomized to three treatment arms : chloroquine plus sulfadoxine-pyrimethamine (CQSP), ammodiaquine plus sulfadoxine-pyrimethamine (AQSP), and ammodiaquine plus artesunate (AQAS). CQSP is the standard treatment and has been widely used while AQSP and AQAS are the new treatments under clinical investigation. Participants were evaluated for malaria infection at baseline (i.e. before treatment), 3, 7, 14, 21 and 28 days after start of treatment by looking at the genotype of the infecting strains to see if treatment had been successful. Baseline covariates include age, gender, parasite density, temperature and genotype.

In this paper, the outcome of interest was infection status 28 days after start of treatment and the analysis is restricted to two treatment arms (CQAS and AQSP). The outcome could be classified as a treatment failure, success or indeterminate. Overally, 25% of 2048 patients had indeterminate outcomes. 33% of the 1018 patients who received the standard of care treatment (CQSP) and 17% of 1030 who received AQSP had indeterminate outcomes respectively. Among patients with observed outcomes, 50% of patients receiving standard of care treatment (CQSP) failed compared to a 18% failure rate among patients receiving the experimental treatment (AQSP).

We estimated the probability of being observed using a logistic regression model with treatment, age, gender, baseline temperature, log baseline parasite density, baseline number of alleles (malaria strains), transmission intensity and time on treatment as predictors. The weights were calculated by taking the inverse of the probability. We estimated the marginal effect of treatment. 95% confidence intervals were estimated from 1000 bootstrap samples of the data. Because missingness was strongly associated with transmission intensity we investigated treatment effects across different transmission intensities.

4.1 Results

Table 4 shows the estimated coefficients, corresponding treatment-specific failure rates and associated 95% confidence intervals. The estimated treatment coefficients for the experimental drug (AQSP) agree very well for all estimators except the two extreme case estimators. However, the complete case estimate for CQSP differs slightly from the MI, IPCW and DR estimators. This difference could be due to the higher proportion of missing outcomes in the CQSP arm (33% compare to 17% in the AQSP arm). Ignoring indeterminate outcomes results in underestimation of the true failure rate by about two percentage points. There are significant differences in failure rates between the two treatments. The experimental treatment (AQSP) is almost three times more effective in preventing malaria reinfection than the standard regimen (CQSP). Using any of the four non-extreme methods would result in similar rate ratios (CC = 2.80, MI = 2.75, IPCW = 2.75, DR = 2.81) but the extreme case methods would result in lower rate ratios (EC-NF = 2.26, EC-F = 2.10)

Table 5 gives the estimated risk of failure for each estimator by transmission intensity. MI,IPCW and DR estimators have similar estimated failure risk at each transmission intensity level for each antimalarial drug. Generally, the risk of treatment failure appear to decrease with increasing transmission intensity, although differences in failure risk are minor between mid-level transmission intensities (2-5).

For CQSP, the DR estimates are closer to the EC-F estimates in low transmission sites than high transmission sites. However, for AQSP, the DR estimates are closer to EC-NF in high transmission sites than in low transmission sites.

5 Discussion

Our analysis demonstrates that in the presence of high degrees of missingness, ignoring the missingness mechanism can result in biased estimates of treatment effects. With a correct model for the probability of missingness, the inverse-probability weighted estimators perform better than the complete case and multiple imputation estimators. The double robust estimator is the most robust of all the estimators discussed.

In malaria studies, transmission intensity significantly confounds treatment effect. Among patients treated with CQSP, missingness rates are higher compared to those treated with AQSP, and the IPCW and DR estimates are much closer to the extreme case estimator in which all mixed genotypes are treated as failures in lower transmission sites whereas the IPCW and DR estimates are closer to the extreme case estimator in which all mixed genotypes are treated as non-failures in high transmission sites. Among AQSP treated patients, in almost all sites except the site with the lowest transmission intensity, the IPCW and DR estimators are closer to estimates where mixed genotypes are assumed to be non-failures. These results seem to suggest that as you move from a poorly effective drug (CQSP) to a more effective drug (AQSP), mixed genotypes are less likely to be recrudesences (old infections). At the same time, as you move from low transmission sites to high transmission sites, mixed genotypes are more likely to be new infections. Thus the size and direction of bias in the extreme case estimators depends on drug as well as transmission intensity.

The estimators we have discussed are all easily implemented using the free R software. Our code is included in the appendix.

Table 1: **Simulation study comparing estimators**

p_a	**Estimator**	$\hat{p_a}$	**BIAS**	**STDBIAS**	**RMSE**
	CC	41.0%	-8.8%	-156.5	0.105
$p_0 = 49.9\%$	MI	50.0%	0.1%	2.4	0.067
	G-Comp	50.0%	0.1%	2.94	0.049
	IPCW	49.6%	-0.2%	-4.1	0.062
	DR	46.5%	-3.4%	-29.8	0.117
	EC-NF	31.7%	-18.2%	-384.5	0.188
	EC-F	54.4%	4.5%	91.3	0.067
	CC	9.0%	-6.6%	-187.2	0.075
$p_1 = 15.7\%$	MI	17.0%	1.3%	16.8	0.079
	G-Comp	17.1%	0.1%	19.0	0.078
	IPCW	15.0%	-0.6%	-6.2	0.101
	DR	13.4%	-2.3%	-26.1	0.090
	EC-NF	4.5%	-11.2%	-613.8	0.075
	EC-F	54.4%	38.8%	783.9	0.391

CC = Complete Case, MI = Multiple Imputation, G-Comp = G-computation, IPCW = Inverse Probability of Censoring Weighted, DR = Double Robust, EC = Extreme Case

Table 2: **Causal Effect: Risk Difference and Relative Risk**

Estimator	Risk Difference $E[Y_0] - E[Y_1]$	Risk Ratio $\frac{E[Y_0]}{E[Y_1]}$
Truth	34.2%	3.18
Complete Case	32.0%	4.56
Multiple Imputation	33.0%	2.94
G-Computation	32.9%	2.92
IPCW	34.6%	3.31
Doubly Robust	33.1%	3.47
EC - Non-failures	27.2%	7.04
EC - Failures	0%	1.00

Table 3: **Robustness to model misspecifications**

Estimator	$E(Y\|A,W)$		$\pi_n(A,W)$		Performance
	bias	std. err	bias	std. err	
MI	-8.6%	5.7%	0.1%	5.3%	poor,good
G-comp	-8.6%	5.5%	0.1%	4.9%	poor,good
IPCW	-0.009%	6.0%	-8.7%	5.6%	good, poor
DR	-0.1%	5.9%	-2.0%	5.8%	good,good
MI	-7.0%	3.5%	0.9%	7.9%	poor,good
G-comp	-6.8%	3.4%	1.0%	7.6%	poor,good
IPCW	-1.5%	8.9%	-6.6%	3.5%	good, poor
DR	-1.7%	7.9%	-3.8%	4.9%	good, fair

Table 4: **Estimated Treatment Coefficients and Failure Rates**

Estimator	**CQSP**	**AQSP**
Complete Case		
Coefficient	-0.003 [-0.153 , 0.140]	-1.523 [-1.748 , -1.306]
Failure Rate	0.499 [0.462 , 0.535]	0.178 [0.130 , 0.238]
Multiple Imputation		
Coefficient	0.112 [-0.028 , 0.256]	-1.552 [-1.753 , -1.313]
Failure Rate	0.528 [0.493 , 0.564]	0.192 [0.144 , 0.258]
G-Computation		
Failure Rate	0.464 [0.493 , 0.564]	0.261 [0.144 , 0.258]
IPCW		
Coefficient	0.125 [-0.008 , 0.270]	-1.553 [-1.781 , -1.323]
Failure Rate	0.531 [0.498 , 0.567]	0.193 [0.143 , 0.259]
Doubly Robust		
Coefficient	0.088 [-0.052 , 0.226]	-1.560 [-1.786 , -1.332]
Failure Rate	0.522 [0.487 , 0.556]	0.186 [0.137 , 0.249]
Extreme Case - non failures		
Coefficient	-0.686 [-0.825 , -0.560]	-1.060 [-1.268 , -0.863]
Failure Rate	0.335 [0.305 , 0.364]	0.148 [0.110 , 0.194]
Extreme Case - failures		
Coefficient	0.681 [0.556 , 0.808]	-1.451 [-1.632 , -0.271]
Failure Rate	0.664 [0.636 , 0.692]	0.316 [0.254 , 0.631]

IPCW = Inverse Probability of Censoring Weighted

Table 5: **Estimated Treatment specific Risk of Failure by transmission intensity**

		*Intensity**					
		1	2	3	4	5	6
CQSP	% missing	25	22	34	26	54	37
	CC	0.708	0.459	0.479	0.469	0.581	0.304
	MI	0.725	0.500	0.524	0.471	0.590	0.378
	G-comp	0.649	0.428	0.454	0.412	0.508	0.334
	IPCW	0.719	0.495	0.534	0.482	0.588	0.396
	DR	0.718	0.491	0.510	0.481	0.583	0.395
	EC-NF	0.532	0.359	0.318	0.349	0.265	0.190
	EC-F	0.780	0.577	0.653	0.605	0.809	0.564
AQSP	% missing	14	7	21	14	27	16
	CC	0.359	0.134	0.144	0.131	0.216	0.083
	MI	0.382	0.138	0.174	0.142	0.209	0.099
	G-comp	0.485	0.223	0.256	0.189	0.312	0.131
	IPCW	0.369	0.144	0.161	0.136	0.222	0.092
	DR	0.367	0.143	0.151	0.135	0.220	0.087
	EC-NF	0.307	0.124	0.113	0.112	0.158	0.069
	EC-F	0.382	0.195	0.327	0.259	0.427	0.231

CC = Complete Case, MI = Multiple Imputation, IPCW = Inverse Probability of Censoring Weighted, DR = Double Robust, EC = Extreme Case
= 1-6 indicates increasing levels of transmission intensity

6 *Appendix*

6.1 R Code

```
# Function to estimate coefficients under varying conditions
# 1. Complete Case estimator - delete all cases with indeterminate outcomes
# 2. Extreme Case estimator - recode all indeterminate as successes delta=0 if
# 3. Extreme Case estimator - recode all indeterminate as failures delta=1 if g
# 4. IPCW estimator - inverse probability of missingness weighted estimating eq
# 5. DR - IPCW estimating equations projected onto the missingness nuisance
# tangent space
# 6. MI - multiply impute missing observations

library(gam)
estimates = function(xdata,ind) {
  attach(xdata)
  wtrt=as.data.frame(cbind(aqsp))
  wcov=cbind(agedich,gender,temp0,logpara,tmn2,tmn3,tmn4,tmn5,tmn6,survival,ban
  n=length(ind)
  glm.0 = glm(delta~aqsp, data=wtrt,family="binomial")

  delta[gmma==0]=0 # missing data
# missingness weights
  glm.gmma = gam(gmma~aqsp+wcov, family=binomial)
  vs.wt = gmma/predict.gam(glm.gmma,type="response")

# observed data
  obs.dg = delta[gmma==1]
  obs.wtrt = as.data.frame(cbind(aqsp[gmma==1]))
  colnames(obs.wtrt)="aqsp"
  obs.wg = wcov[gmma==1,]

# Fitting models
  glm.1 = glm(obs.dg~aqsp, data=obs.wtrt, family=binomial)
  glm.2 = glm(delta~aqsp, data=wtrt,family=binomial)
  delta[gmma==0]=1
  glm.3 = glm(delta~aqsp, data=wtrt, family=binomial)
```

```
  delta[gmma==0]=0
  glm.4 = glm(delta~aqsp, data=wtrt, family=binomial, weights=vs.wt)

# Double Robust estimate - the idea is that the IPCW only uses observed data.
# But we could gain efficiency by using extra data from that with missing outco
  piw = predict(glm.gmma, type="response")

# Estimate E[eps2|X*] using gam
  initial.ipcw.estimate = glm(delta~aqsp, data=wtrt, family=binomial(), weights
  residuals.init.ipcw = delta - initial.ipcw.estimate$fitted
  res.sqd = residuals.init.ipcw^2
  res.gam = gam(res.sqd~aqsp, data=wtrt, weights=vs.wt)
  predict.expect.res = predict.gam(res.gam, type="response")

# derivative matrix mdot = fn(W)*[1 a W]
  xx.design.matrix = cbind(1,as.matrix(wtrt))
  fntw = exp(-1*predict.glm(initial.ipcw.estimate))/(1+exp(-1*
  predict.glm(initial.ipcw.estimate)))^2
  mdot = fntw*xx.design.matrix
# h(X*) = mdot*1/E[eps2|X*]
  hxstar = diag(1/predict.expect.res) %*% mdot

# Model E[gamma*h(X*)*eps(beta)/pi(w) |a w] each component
# this is the projection on to the nuisance tangent space
  hxstar[gmma==0]=0
  h1.proj = hxstar[,1]*residuals.init.ipcw*vs.wt
  h2.proj = hxstar[,2]*residuals.init.ipcw*vs.wt
#  h3.proj = hxstar[,3]*residuals.init.ipcw*vs.wt

  h1.proj.gam = gam(h1.proj~aqsp+wcov, data=wtrt)
  h2.proj.gam = gam(h2.proj~aqsp+wcov, data=wtrt)
#  h3.proj.gam = gam(h3.proj~aqsp+wcov, data=wtrt)

  h1.pred.gam=numeric(length(h1.proj))
  h2.pred.gam=numeric(length(h1.proj))
#  h3.pred.gam=numeric(length(h1.proj))

  h1.pred.gam.temp = predict.gam(h1.proj.gam)
```

```
  h2.pred.gam.temp = predict.gam(h2.proj.gam)
#  h3.pred.gam.temp = predict.gam(h3.proj.gam)

  h1.pred.gam[as.numeric(names(h1.pred.gam.temp))] = h1.pred.gam.temp
  h2.pred.gam[as.numeric(names(h2.pred.gam.temp))] = h2.pred.gam.temp
#  h3.pred.gam[as.numeric(names(h3.pred.gam.temp))] = h3.pred.gam.temp
  expect.hst.eps = cbind(h1.pred.gam,h2.pred.gam)

#Objective function
  dr.objfn = function(theta) {
    resa = as.vector(((delta - (1/(1+exp(-1*xx.design.matrix %*%
  theta))))*gmma)/piw)
    ic = resa * hxstar - ((gmma-piw)/piw)*expect.hst.eps
    obj = sum(sqrt(sum(apply(ic,2,mean)^2)))
    obj
  }

# get parameter estimates using optim
  vcoeff=as.vector(initial.ipcw.estimate$coeff)
  beta.opt= optim(vcoeff,dr.objfn,method="BFGS",
  control=list(abstol=0.000001, maxit=500))
  dr.coeff = beta.opt$par

  # Imputation


  # Model P[delta = 1 | A, W] conditional distribution of failure using observe
  obs.wall = data.frame(obs.wtrt, obs.wg)
  wall = data.frame(wtrt, wcov)
  glm.delta =  glm(obs.dg~aqsp+agedich+gender+temp0+logpara+survival+bands0+
  tmn2+tmn3+tmn4+tmn5+tmn6, data=obs.wall, family=binomial())

  # Estimated probabibility of failure
  prob.delta.hat = predict.glm(glm.delta, newdata=wall, type="response")

  imp.coeff.out=NULL

  # generate new deltas repeatedly, estimate the coefficients, average
```

```
  for(j in 1:5) {
    new.d = rbinom(length(prob.delta.hat),1,prob.delta.hat)
    delta.n = delta
    delta.n[gmma==0] = new.d[gmma==0]
    imp.coeff.out = rbind(imp.coeff.out,
    coefficients(glm(delta.n~aqsp, data=wtrt,
    family=binomial())))
  }
  imp.coeff.mean = apply(imp.coeff.out,2,mean)
  aqsp=c(0,1)
#  aqas=c(0,0,1)
  nmd=data.frame(cbind(aqsp))
  p.0 = predict(glm.0, newdata=nmd, type="response")
  p.1 = predict(glm.1, newdata=nmd, type="response")
  p.2 = predict(glm.2, newdata=nmd, type="response")
  p.3 = predict(glm.3, newdata=nmd, type="response")
  p.4 = predict(glm.4, newdata=nmd, type="response")
  p.imp = c(1/(1+exp(-(imp.coeff.mean[1]))),1/(1+exp(-(imp.coeff.mean[1]+imp.co
  p.dr = c(1/(1+exp(-(dr.coeff[1]))),1/(1+exp(-(dr.coeff[1]+dr.coeff[2]))))
# return coefficients from all the procedures
  txcoeff=as.vector(c(coefficients(glm.1),coefficients(glm.2), coefficients(glm
  imp.coeff.mean,coefficients(glm.4),dr.coeff,p.1,p.2,p.3,p.imp,p.4,p.dr))
return(txcoeff)
}
ind=mal2$id
txcoeff = estimates(mal2,ind)

# Bias and Variance estimation by Bootstrap
fname="c:/malaria/boot.out"
    for (bs in 1:1000) {
      boot.dat=mal2[sample(row.names(mal2), replace=TRUE),]
      boot.beta.a = estimates(boot.dat,1:length(boot.dat))[1:12]
      if (bs==1) {
        cat(file=fname,boot.beta.a[1:12],"\n")
      } else {
        cat(file=fname,boot.beta.a[1:12],"\n",append=TRUE)
      }
      cat(bs,"\n")
```

```
detach()
}

boot.beta = read.table("c:/malaria/boot.out",sep=" ")[1:12]
bt.st = apply(boot.beta,2,sort)
ci.95 = bt.st[c(25,975),]
cat(file="mal-analysis.out","\n\n", append=TRUE)
cat(file="mal-analysis.out","95% Confidence Interval","\n", append=TRUE)
cat(file="mal-analysis.out",ci.95,"\n", append=TRUE)
```

References

[1] Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[2] Richard D. Gill, Mark J. van der Laan, and James M. Robins. Coarsening at random: Characterizations, conjectures and counter-examples. In *Proceedings of the First Seattle Symposium in Biostatisticss, 1995, D.Y. Yin and T.R. Fleming (eds), Lecture Notes in Statistics*, pages 255–294. Springer, New York, 1997.

[3] Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19:2244–2253, 1991.

[4] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002.

[5] Fabrizia Mealli and Donald B. Rubin. Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Services & Outcomes Research Methodogy*, 3:225–232, 2002.

[6] James Robins and Andrea Rotnitzky. Estimation of treatment effects in randomised trials with non-compliance and a dichotomous outcome using structural mean models. *Biometrika*, 91(4):763–783, 2004.

[7] James M. Robins. Comment on "Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard"(p1641-1661). *Statistics in Medicine*, 21(12):1663–1680, 2002.

[8] Andrea Rotnitzky, James M. Robins, and Daniel O. Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93:1321–1339, 1998.

[9] Donald B. Rubin. Inference and missing data. *Biometrika*, 63:581–590, 1976.

[10] Donald B. Rubin. Reply to comments on "Multiple imputation after 18+ years". *Journal of the American Statistical Association*, 91:515–517, 1996.

[11] Daniel O. Scharfstein and Rafael A. Irizarry. Generalized additive selection models for the analysis of Studies with potentially nonignorable missing outcome data. *Biometrics*, 59(3):601–613, 2003.

[12] Daniel O. Scharfstein and James M. Robins. Estimation of the failure time distribution in the presence of informative censoring. *Biometrika*, 89(3):617–634, 2002.

[13] M. Slater, M. Kiggundu, C. Dokomajilar, MR. Kamya, N. Bakyaita, A. Talisuna, PJ. Rosenthal, and G. Dorsey. Distinguishing recrudescences from new infections in antimalarial clinical trials: major impact of interpretation of genotyping results on estimates of drug efficacy. *American Journal of Tropical Medicine and Hygiene*, 73(2):256–262, 2005.

[14] G. Snounou and HP. Beck. The use of pcr-genotyping in the assessment of recrudesence or reinfection after antimalarial drug treatment. *Parasitology Today*, 14:462–467, 1998.

[15] Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, 2002.

[16] Angela M Wood, Ian R White, and Simon G Thompson. Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, 1:368–376, 2004.