Research Article

Cletus Kwa Kum*, Daniel Thorburn, Gebrenegus Ghilagaber, Pedro Gil,
and Anders Björkman

On the Effects of Malaria Treatment on Parasite Drug Resistance – Probability Modelling of Genotyped Malaria Infections

Abstract: We compare the frequency of resistant genes of malaria parasites before treatment and at first malaria incidence after treatment. The data come from a clinical trial at two health facilities in Tanzania and concerns single nucleotide polymorphisms (SNPs) at three positions believed to be related to resistance to malaria treatment. A problem is that mixed infections are common, which both obscures the underlying frequency of alleles at each locus as well as the associations between loci in samples where alleles are mixed. We use combinatorics and quite involved probability methods to handle multiple infections and multiple haplotypes. The infection with the different haplotypes seemed to be independent of each other. We showed that at two of the three studied SNPs, the proportion of resistant genes had increased after treatment with sulfadoxine–pyrimethamine alone but when treated in combination with artesunate, no effect was noticed. First recurrences of malaria associated more with sulfadoxine–pyrimethamine alone as treatment than when in combination with artesunate. We also found that the recruited children had two different ongoing malaria infections where the parasites had different gene types.

Keywords: genotyped malaria, efficacy, combinatorics, parametric model, haplotype, single nucleotide polymorphism (SNP)

*Corresponding author: Cletus Kwa Kum, Department of Statistics, Stockholm University, Stockholm, Sweden; Department of Mathematics/Computer Science, Faculty of Science, University of Dschang, Dschang, Cameroon, E-mail: cletus.kwa.kwa@stat.su.se, kumkwah@yahoo.co.in

Daniel Thorburn: E-mail: Daniel.Thorburn@stat.su.se, Gebrenegus Ghilagaber: E-mail: Gebrenegus.Ghilagaber@stat.su.se, Department of Statistics, Stockholm University, Stockholm, Sweden

Pedro Gil: E-mail: pedgil01@gmail.com, Anders Björkman: E-mail: anders.bjorkman@karolinska.se, Division of Infectious Diseases, Karolinska University Hospital, Stockholm, Sweden

1 Introduction

1.1 Background

Malaria is caused by parasite of the genus *Plasmodium*. The disease is one of the leading causes of death in most tropical parts of the world including Africa. *Plasmodium falciparum* is the most fatal species. Several studies and clinical trials have been made to study malaria and the efficacies of different therapies, particularly in the context of the capacity of this organism to develop drug resistance. Our goal in this article is to use data of a study on artemisinin-based combination therapies (ACT) to study the proportion of parasites believed to be drug resistant before treatment and in observed first recurrence of the disease afterwards. In other words, is there a selection enhancing the frequency of parasites present in recurrent infections which carry gene variants associated with drug resistance?

The problem of estimating the proportions of different haplotypes from blood samples with multiple genes has usually been approached using simple methods like neglecting all mixed infections with multiple genotypes or counting the multiple genes as resistant. Wigger et al. [1] uses Markov chain Monte Carlo (MCMC) methods where one step simulates the true state. Hastings and Smith [2], present a computer package for the calculations. Other methods have been reported by Hill and Babiker [3] and Kessner et al. [4]. These papers mostly deal with one population at one time point (at baseline or at recurrence separately). When comparisons are made, they are made later on the independent estimates. In this article, we develop models for several related populations jointly, where only a few parameters measure changes.

Another problem is estimating the number of malaria infections using blood samples data. This has been previously treated by Ross et al. [5] among others. We show how such estimates can be obtained as a byproduct of our estimates of the proportions assuming independence between infections.

1.2 Description of the clinical study

The Malaria Research Unit of the Clinical Epidemiology Unit of Karolinska Institutet, Sweden, conducted a clinical trial in 2004 to compare the efficacies of two therapies in treating children under five years infected with uncomplicated *Plasmodium falciparum* malaria parasites. A description of a similar study and full details regarding the entry requirements and the conduct of the trial can be found in Mårtensson et al. [6]. The study was undertaken at two health facilities in Tanzania – Uzini and Kondé with 206 and 178 uncomplicated malaria patients, respectively. At entry to the study, blood samples were taken from all patients. The children were randomly allocated into two treatment arms. In one arm, the children were treated with only sulfadoxine–pyrimethamine (SP) and in the other arm, they were treated with artesunate plus sulfadoxine–pyrimethamine (ASP). The children were followed for 84 days after treatment and retested after 7, 21, 28, 42, 56 or 84 days. If the child had a recurrent malaria infection, the parasites were genotyped. For each child, we got the gene sequence at baseline and at the first recurrence of malaria, which could be due to a reinfection or a recrudescence. There was no possibility to separate these two reasons for parasite recurrence with certainty.

During the trial period, some children were free from malaria during the whole follow-up periods. For 109 children from Kondé and 88 children from Uzini, we had only blood samples at baseline. Using Bayesian survival models, the combination treatment with ASP was stronger in averting recurrent infection than SP alone [7]. The difference was quite clear during the first weeks immediately after treatment, probably reflecting treatment failures (recrudescences), whereas that difference had almost disappeared after three months. By that time, recurrent infections were probably only represented by new infections (reinfections) which occurred supposedly similarly in both treatment arms. The parasites in the blood samples were analysed and the single nucleotide polymorphisms (SNPs) at three positions in the *pfdhfr* gene were determined. The three positions (*pfdhfr* 51, 59 and 108) could be defined as either resistant (R) or sensitive (S). These positions were known to be important for the resistance of the parasite to SP. If both parasites with R and S SNPs were present, this was denoted by the letter M. Each blood sample was classified for its parasites *pfdhfr* characteristics by three letters from SSS to MMM, denoting the status of each one of the individual SNPs. For example, RSM means that the child had only parasites with resistant SNPs at the first position and only sensitive SNPs at the second position, but there were both parasites with resistant and sensitive SNPs at the third position. In the study, many other important aspects and properties were studied, but we will only analyse this part of data.

Our main interest is the development of the proportion of resistant SNPs. It is common practice to analyse the frequencies of individual SNPs and classify mixed alleles as resistant [2]. Results from analyses with such assumption are usually biased. In this article, we will develop a method to use the information in an adequate way and more importantly, information from those patients carrying the M (mixed R and S) SNPs. Our hypothesis is that, there are relatively more resistant SNPs at the first recurrence of malaria if we correct for the fact that the children at baseline were infected by multiple types of malaria parasites simultaneously. The

data are given in Table 6. Blood sample phenotypes observed in such a study do not necessarily describe the original combinations completely. For instance, the combinations $RRR + SRS$ and $RRS + SRR$ will both yield the classification MRM, as will $RRR + RRS + SRR$, $RRR + RRS + SRS$ and three other combinations. All of such possible combinations are taken into account and a complete description of possible types is presented in Table 7. The fact that the true genotype proportions are not observed when multiple genes occur call for rather complicated calculations; since the amount of infection was higher at baseline compared to at first recurrence of malaria where more genes of multiple types were observed.

1.3 Short overview of the article

In the next section, we derive models for data considering only one period (baseline or at first recurrence of malaria). We study what observations to expect if the proportion of different parasites is known, taking into account that the children may have multiple infections. The frequencies of the 27 observed types and their probabilities are derived from the observed population frequencies. The population parameters are estimated using maximum likelihood (ML) methods. There is shown to be a positive relation between resistant genes at the three SNPs. For example, the presence of a resistant gene at one loci increases the probability that the parasite also has resistant genes at the other positions. We study the number of different malaria haplotypes an average child has. In Section 3, we combine the two periods and derive corresponding expressions, modelling the treatment effects with common parameters. Finally, we combine the data from the two health facilities, Kondé and Uzini. The article ends with a discussion.

2 One time-point probability models

2.1 The saturated model

For completeness we start with a saturated model. It will be used as an alternative hypothesis when testing different reduced models. There were $3^3 = 27$ different possible malaria statuses at baseline and only 28 at the first reappearance of malaria, since no recurrence was possible at baseline. Let these genotypes be represented by IJK , where $I = M, R$ or S , $J = M, R$ or S and $K = M, R$ or S . Further, let n_{IJK} be the corresponding number of patients with this infection. Then the probability of an infection IJK can be estimated by the corresponding relative frequency, that is,

$$\pi_{IJK}^* = \frac{n_{IJK}}{N}, \quad (1)$$

where N is the total number of observed patients with infection. These estimates are presented in Table 6.

2.2 Relation between true haplotypes and the observations with multiple genes

2.2.1 Model derivation

The parasites can be classified into eight haplotypes.

$$XYZ \in \{RRR, RRS, RSR, SRR, RSS, SRS, SSR, SSS\}.$$

In this article, we will use the letters IJK when we classify patients/observations into R, S or M , but XYZ when classifying parasites with only R or S . Let the probability of a susceptible child to be infected by type XYZ at the time point be

$$p_{XYZ} = 1 - q_{XYZ}, \quad (2)$$

where X, Y and Z may be R or S. Now, assume that these eight haplotypes infect children independently of each other. In that case the probability that a child stays healthy, that is, is free from all eight possible parasite types is

$$\pi_{(H)} = \prod_{XYZ}^8 q_{XYZ}, \quad (3)$$

where the product is over the eight possible parasite types.

The probability of a child being infected with exactly one of these types can also be calculated. For example, the probability of an RRR infection only is given by

$$\pi_{RRR} = \left(\frac{p_{RRR}}{q_{RRR}} \right) \prod_{XYZ} q_{XYZ}. \quad (4)$$

We next turn to the event of being infected with several types. The probability of an MRR outcome is

$$\pi_{MRR} = \left(\frac{p_{RRR}}{q_{RRR}} \times \frac{p_{SRR}}{q_{SRR}} \right) \prod_{XYZ} q_{XYZ}. \quad (5)$$

This is because MRR corresponds to an infection with both RRR and SRR but no other. All probabilities of infections with genotypes with only one M can be obtained in this way. It is also relatively easy to calculate the probability of genotypes with two M's.

There are 12 possible classifications with one M and six classifications with two M. Each of this six cases can be obtained from seven different combinations of infection types (see Table 7). The gene classification MMM is more complicated, since it can be obtained in 193 ways. The total number of ways to combine the eight types is $2^8 = 256$ (255 of these, give malaria and one is free from infection)

The probability of MMM can be obtained by summing 193 terms but it is probably simpler to subtract the sum of all the other probabilities from 1.

$$\pi_{MMM} = 1 - \sum_{IJK} \pi_{IJK}, \text{ where } IJK \neq MMM. \quad (6)$$

However, we cannot estimate all these values. In the clinical trials, only young parasitaemic children were recruited. Therefore (at baseline) we only observe proportions conditional on the event of having malaria. Thus the probability of an observation classified as IJK is

$$\pi'_{IJK} = \frac{\pi_{IJK}}{\left(1 - \prod_{XYZ} q_{XYZ} \right)}, \quad (7)$$

where $I = M, R \text{ or } S, J = M, R \text{ or } S$ and $K = M, R \text{ or } S$.

2.2.2 Statistical estimation

The likelihood function is

$$\mathcal{L}(\mathbf{p}) = \prod_{IJK} \left[\pi_{IJK} / \left(1 - \prod_{XYZ} q_{XYZ} \right) \right]^{n_{IJK}} \quad (8)$$

where $\mathbf{p}$ is the vector of the eight probabilities. We maximise this function to obtain estimates of the probabilities of XYZ infections by the maximum likelihood method. A programme code was written and the optimisation technique used was the Nelder and Mead [8] method in R. This technique does not require

derivatives which makes it suitable for optimisation of non-smooth functions. It often shows rapid improvements with a relatively small number of iterations.

We use this model for the baseline data and also for the data at first observed recurrence of malaria, at each of the two health centres. Even though some healthy children were observed, the probability of staying healthy can not be estimated using eq. (3), due to the sampling design. The children were tested seven times with an average interval of 14 days, which means that all observed first recurrence cases occurred about two weeks after being found healthy in the previous test. However, those being healthy at the end of the study had been so for almost three months. Therefore, the described modelling procedure for first recurrences will not apply to no recurrences.

The optimisation procedure employed produced maximum likelihood (ML) estimates for p_{XYZ} , which are presented in Tables 1 and 2, each for Kondé and Uzini, respectively. The decrease in $2 \times \log\text{likelihood}$ from the saturated model was 20.5 for Kondé at baseline and 26.1 in Uzini at baseline. Since the decrease in the number of parameters is $27 - 8 = 19$, we may accept that this model holds for both health centres. Thus, we can safely assume that the eight infection types infect independently of each other. The corresponding test values at first recurrence of malaria were 6.5 and 36.8, respectively. It is usual to compare these values to a percentile of the asymptotic χ^2 -distribution. In these two cases, the asymptotic distribution has not been reached but the figures nevertheless indicate that the model with independent parasite types seems to be reasonable also for data at first recurrence.

If we assume that the parasite types infect independently of each other, we may estimate the proportion of the eight haplotypes in the surrounding environment. We compute the eight frequencies of the different parasite types as,

$$f_{XYZ} = \lambda_{XYZ} / \sum_{UVW}^8 \lambda_{UVW}, \quad (9)$$

where $\lambda_{XYZ} = -\ln(q_{XYZ})$. These estimated frequencies, f_{XYZ}^* , are also found in Table 1, where no distinction is made between the treatments, and in Table 2, where the distinction is made.

One may note two things. The parasite type RRR with resistant genes at all three positions is the most common type (see Table 1). The combination SRR is rare. The appearances of the resistant genes at the three positions do not seem to be independent. The proportion of parasites with exactly two sensitive genes seems to be too low and the proportion of parasites with three sensitive genes seems to be too high for independence to hold. This will be checked by a formal test in the next section.

In Table 2, the difference between ASP and SP are just due to chance, since these data were observed before treatment. So treatment could not have influenced the parameters. Meanwhile at first recurrence of malaria, one observes that while there was a relative drop in the proportion of RRR at first recurrence of

Table 1 Estimated haplotype frequencies: no distinction between treatments

Haplotype	KONDE				UZINI			
	Baseline		First recurrence		Baseline		First recurrence	
	p_{XYZ}^*	f_{XYZ}^*	p_{XYZ}^*	f_{XYZ}^*	p_{XYZ}^*	f_{XYZ}^*	p_{XYZ}^*	f_{XYZ}^*
RRR	0.61	0.47	0.43	0.70	0.29	0.42	0.29	0.55
RRS	0.32	0.20	0.06	0.07	0.17	0.23	0.09	0.16
RSR	0.32	0.19	0.14	0.19	0.12	0.16	0.09	0.14
SRR	0.03	0.01	0.00	0.00	0.00	0.00	0.02	0.03
RSS	0.03	0.02	0.03	0.03	0.01	0.01	0.02	0.03
SRS	0.02	0.01	0.00	0.00	0.02	0.02	0.01	0.02
SSR	0.01	0.003	0.00	0.00	0.003	0.004	0.005	0.01
SSS	0.16	0.09	0.00	0.00	0.12	0.15	0.04	0.07

Table 2 Estimated haplotype frequencies: distinction between treatments

Haplotype	KONDE												UZINI			
	ASP						SP						ASP			
	Baseline			First recurrence			Baseline			First recurrence			Baseline		First recurrence	
	p_{xyz}^*	f_{xyz}^*	p_{xyz}	p_{xyz}^*	f_{xyz}^*	p_{xyz}	p_{xyz}^*	f_{xyz}^*	p_{xyz}	p_{xyz}^*	f_{xyz}^*	p_{xyz}	p_{xyz}^*	f_{xyz}^*	p_{xyz}	f_{xyz}^*
RRR	0.584	0.437	0.385	0.465	0.511	0.591	0.630	0.636	0.630	0.511	0.591	0.630	0.278	0.381	0.301	0.450
RRS	0.355	0.219	0.073	0.072	0.176	0.065	0.290	0.063	0.217	0.287	0.088	0.223	0.217	0.287	0.147	0.200
RSR	0.304	0.180	0.332	0.385	0.203	0.102	0.326	0.101	0.112	0.147	0.050	0.123	0.112	0.147	0.119	0.160
SRR	0.028	0.014	0.000	0.000	0.014	0.000	0.027	0.000	0.000	0.000	0.008	0.019	0.000	0.000	0.000	0.000
RSS	0.041	0.021	0.078	0.077	0.033	0.000	0.026	0.000	0.009	0.011	0.023	0.056	0.010	0.013	0.010	0.013
SRS	0.045	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.014	0.016	0.000	0.000	0.018	0.023	0.018	0.023
SSR	0.008	0.004	0.000	0.000	0.004	0.000	0.007	0.000	0.000	0.000	0.009	0.022	0.006	0.008	0.006	0.008
SSS	0.187	0.103	0.000	0.000	0.079	0.000	0.142	0.000	0.126	0.158	0.049	0.121	0.110	0.147	0.110	0.147

disease for children treated with ASP, there was a significant increase in the proportion of RRR for the sick children treated with SP.

2.3 Number of infections

It is interesting to note that the average patient at baseline had many types of infections. The sum of the p_{XYZ}^* column, in Table 1, estimates the average number of different haplotypes of a child. But since our models hold only for sick children, we have to divide by the probability of being healthy given by, $1 - \prod_{XYZ}^8 q_{XYZ}$. For Kondé at baseline the average number of types for a sick child is thus $1.50/0.86 = 1.74$ (see Table 1). The other estimates where distinction was made between treatments are shown in Table 2.

Those having only one type may have been infected many times but with the same type. If we assume that the number of times a child is infected follows a Poisson distribution, we know that $\Pr(0) = \exp(-\lambda)$, where λ is the parameter or the mean. We recall that p_{XYZ} is the probability of getting infected at least once with haplotype XYZ. Since we assume a Poisson distribution, then λ_{XYZ} equals the expected number of times a person is infected with XYZ. Conditioning on the fact that we are only studying persons who are infected by at least one type gives $\lambda_{XYZ}/(1 - p_{XYZ}) = E(\text{number of times a person, known to be infected given that he is infected at least once})$. A summation over XYZ gives the expected number of times a person is infected given that he is infected at least once.

Thus the expected number of infection times can be estimated by

$$-\sum_{XYZ} \ln(q_{XYZ}) / \left(1 - \prod_{XYZ}^8 q_{XYZ}\right) \quad (10)$$

These figures can be found in Tables 3 and 4.

From Table 3, it is seen that the amount of infection at baseline was much higher at Kondé compared to Uzini. The same pattern follows at first recurrence of malaria for both treatments, shown in Table 4. The

Table 3 Average number of haplotypes and expected number of times patients were infected: no distinction between ASP and SP

	KONDE		UZINI	
	Baseline	First recurrence	Baseline	First recurrence
Average no. of haplotypes	1.74	1.19	1.31	1.21
Expected no. of infection times	2.29	1.46	1.46	1.35

Table 4 Average number of haplotypes and expected number of times patients were infected: distinction between ASP and SP

	KONDE				UZINI			
	ASP		SP		ASP		SP	
	Baseline	First recurrence	Baseline	First recurrence	Baseline	First recurrence	Baseline	First recurrence
Average no. of haplotypes	1.79	1.34	1.69	1.15	1.33	1.15	1.30	1.25
Expected no. of infection times	2.31	1.61	2.27	1.63	1.49	1.22	1.47	1.50

estimated number of infections was still higher at Kondé than at Uzini. The smaller differences in the numbers of haplotypes may be explained by the fact that four types (i.e. S at first position, SRR, SRS, SSR and SSS) had disappeared at Kondé at the time of first recurrence.

2.4 A model assuming independence between gene positions

If the population of parasites was well mixed for a long time, the occurrence of an R or S should be independent at all positions or the frequency f_{XYZ} should be possible to factorize.

$$f_{XYZ} = f_{X..} \times f_{.Y.} \times f_{..Z} = \lambda_{X..} \lambda_{.Y.} \lambda_{..Z} / \sum_{UVW} \lambda_{UVW}, \quad (11)$$

Since p_{XYZ} is fairly small, $\lambda_{XYZ} = -\ln(1 - p_{XYZ}) \approx p_{XYZ}$. Thus factorizing f_{XYZ} is approximately the same as factorizing p_{XYZ} , but we may include a normalizing constant, α . Thus, we tested the model

$$p_{XYZ} = \alpha \times p_{X..} \times p_{.Y.} \times p_{..Z} \quad (12)$$

This model with an α is more robust to errors in estimation of the probability of not having malaria, since an error in that probability will be absorbed by the constant α .

The tests of this four parameter model against the eight parameter model gave the deviances 65.8 (Kondé, baseline), 140.3 (Uzini, baseline), 0.1 (Kondé, at first recurrence) and 29.3 (Uzini, at first recurrence). Three of them reject the hypothesis of independence and the remaining model with test quantity 0.1 builds on so many 0-values in Table 1, and a χ^2 test can be considered meaningless. We thus conclude that occurrences of the three genes are not independent. It may be noted that the proportion of RRR and SSS haplotypes seem to be over-represented.

3 Two time-points probability models

3.1 The combined model

In Section 2, we showed that one may accept the model where the observed ratios are explained by only eight proportions. We will take this as proved. We will now look at baseline and at first recurrence simultaneously. There may be some relation between the two time points. In that case, it may be possible to describe the proportions at first recurrence of the disease with the eight baseline parameters and a few more parameters describing the differences. We will always use likelihood ratios to test whether different models are true. In all tests, we compare the deviance with the asymptotic χ^2 distribution. Sometimes the asymptotic distribution is not reached, but in those cases the conclusion were often quite clear anyway. Details of model selection and deviance methods can be found in Burnham and Anderson [9], Cox [10] and Pawitan [11].

Denoting the parameters at baseline with an extra index b , and at first recurrence of malaria by an extra index r , we get as a starting point the model

$$p_{bXYZ} = 1 - \exp(-\lambda_{bXYZ})$$

$$p_{rXYZ} = 1 - \exp(-\lambda_{rXYZ})$$

Can this be explained by fewer parameters?

3.2 A model with varying amount of infection

It is obvious from Table 1 that there are more infections at baseline than at the first recurrence. One possible reason is that a person, who is untreated for a long period, may have had time to be infected more than once. The longer the patients are exposed, the greater the amount of infection. During the follow-up period, a child will on the average be checked at two weeks intervals. We consider a simple model in which the only difference between baseline and at the first recurrence is that total amount of infection t is smaller at this first appearance of malaria. We model this by letting $\lambda_{rXYZ} = t \times \lambda_{bXYZ}$. This means

$$p_{rXYZ} = 1 - \exp(-t \times \lambda_{bXYZ}) \quad (13)$$

With a slight misuse of words, one might say that if the exposure time at baseline is set to 1, the exposure is only t at the first recurrence (reinfection) with this model.

If this is tested against the model in the previous section, the increase in $2 \times \log\text{likelihood ratio}$ is 16.5 for Kondé and 20.2 for Uzini. With 7 degrees of freedom (d.f.), we must reject this model. We thus look at a model where the decrease is different for different genes.

3.3 A model with varying amounts of infection and differences between gene positions

In Section 3.2, we saw that all eight types of infections do not decrease equally much. In our next model, we will try to model the differences by less than eight parameters. We consider two models, one additive and one multiplicative. These models describe the hypothesis that the proportion of sensitive genes decreases between baseline and at first recurrence of malaria.

$$p_{bXYZ} = 1 - \exp(-t(\alpha_1^{I_1} \times \alpha_2^{I_2} \times \alpha_3^{I_3} \times \lambda_{bXYZ})) \quad (14)$$

$$p_{bXYZ} = 1 - \exp(-t(\alpha_1 I_1 + \alpha_2 I_2 + \alpha_3 I_3 + \lambda_{bXYZ})) \quad (15)$$

where

$$\begin{cases} I_1 = 1 & \text{if } X \text{ in the genotype } (XYZ) = S \text{ and } 0 \text{ otherwise} \\ I_2 = 1 & \text{if } Y \text{ in the genotype } (XYZ) = S \text{ and } 0 \text{ otherwise} \\ I_3 = 1 & \text{if } Z \text{ in the genotype } (XYZ) = S \text{ and } 0 \text{ otherwise} \end{cases}$$

In both models, the parameters α_1 , α_2 and α_3 measure the effect of the first, second and third gene positions, respectively, being an S (and not an R). If $\alpha_i = 1$ for the first model, equation (14), there is no difference between R and S at the corresponding position. If $\alpha_i = 0$ for model 15, then there is no difference between R and S at the corresponding position. Higher values mean an increase in the proportion of haplotypes with R –genes.

Model (14) fits the data quite well. The increase in $2 \times \log\text{likelihood ratio}$ was 1.8 for Kondé and 9.4 for Uzini which should be compared with a χ^2 distribution with 4 d.f. (However, we note that since there were no S genes at the first position at the first recurrence, the ML estimate is $\alpha_1 = 0$ and the conditions for an asymptotic $\chi^2(4)$ – distribution does not hold.) The second model, that is model (15), did not fit the data as well as the multiplicative model given by eq. (14). The test quantities were 15.0 and 12.1, respectively. We thus settled for the multiplicative model. The estimated values for the four parameters are given in Table 5. The conclusion at Kondé is that there were no S genes left at the first position at the first recurrence of malaria. The proportion of R genes at the last position was much higher at recurrence and somewhat higher, but not significant, at the second position. (The value of t shows a clear decrease in the amount of infection at Kondé, but that depends more on the fact that the children were more infected at the outset in Kondé.) At

Table 5 Measure of the relative increase in haplotypes with S genes at different positions

Parameter	Kondé	Uzini	All data	ASP	SP
t_K	0.56	–	0.48	0.579	0.439
t_U	–	1.07	1.12	0.512	1.915
α_1	0.00	1.29	0.98	0.605	1.073
α_2	0.83	0.69	0.75	1.464	0.486
α_3	0.37	0.49	0.47	0.882	0.304

Uzini, the decrease in S genes was significant only at the last position. (It was even a non-significant increase at the first position.)

3.4 A model combining both health centres

In the foregoing section, we ended up in the same model at both health centres but with different parameter values. A natural question is whether the differences were due to chance. We thus start this section by testing whether the three α parameters are the same in Kondé and Uzini, that is, if the true relative decrease in the proportion of S genes at the three positions can be the same at both locations.

$$p_{rKXYZ} = 1 - \exp(-t_K(\alpha_1^{I_1} \times \alpha_2^{I_2} \times \alpha_3^{I_3} \times \lambda_{bKXYZ})) \text{ at Kondé} \quad (16)$$

$$p_{rUXYZ} = 1 - \exp(-t_U(\alpha_1^{I_1} \times \alpha_2^{I_2} \times \alpha_3^{I_3} \times \lambda_{bUXYZ})) \text{ at Uzini} \quad (17)$$

The resulting test quantity was 7.4 with 3 d.f. Therefore, we cannot reject the hypothesis and we may accept that the same model can be used at both centres. We also tested whether the exposure parameters, t , were the same but as could be expected this was rejected; since the amount of infection at baseline was different (test quantity 42.4, d.f. = 1).

The measures of the relative increase of haplotypes with S genes at different positions are also presented in Table 5. These estimates are obtained using the multiplicative models (14), (16) and (17). We equally have estimates for Kondé and Uzini for both treatments combined, ASP and SP combined for the two centres.

All ASP α 's are around 1. There is no real decrease in the proportion of sensitive genes for infections treated with ASP, but for the last positions. The decrease is clear for those treated with SP. Where we earlier had a decrease by the factors 0.75 and 0.47 (see Table 5), it is now only 0.49 and 0.30 for SP. The decrease is most obvious in Uzini, where there were more infections (t value = 1.91).

4 Results and discussion

Our main question was to study the relative proportion of resistant genes before and at first recurrence of malaria since the start of treatment. We developed methods and used them in the estimation of population frequencies of haplotypes of three SNPs using observed frequencies in samples. These SNPs were R(resistant), S(sensitive) and M(mixed R and S) SNPs. We divided all blood samples with multiple genes (M) into separate infection combinations. Since almost half of the samples had multiple genes which could correspond to 255 possible infection combinations, this meant use of much combinatorics. An analysis of only those samples without multiple genes would mean that only half the data set was used and also that the small haplotype proportions would be under-estimated, since they were most likely to be observed together with other more frequent haplotypes and the most frequently over-estimated. For instance f_{RRR} would have been estimated by 0.58 and not 0.47 as in Table 1. Less infection may lead to fewer multiple genes observed and even an increased number of S genes. In this case, our careful modelling of the

populations at two time points will not work. Let us just give a simple example with one gene, where the infecting population consists of $2/3$ R and $1/3$ S. A population with at average has three infections; the sick population will consist of 58% M, 33% R and 9% S. With only one half infection per person, the corresponding figures are 11% M, 61% R and 28% S. Just by reducing the total number of infections, the ratio between R and S drops considerably (by 70%). It is obvious that these cannot be used when deciding whether the proportion of S has decreased. Another choice might be to include all data with only one multiple genes dividing them between the components. This would be better only if we leave out those observations where the combination was not certain. But still, one quarter of the Kondé data would be left out and the errors would be in the same direction.

We could not reject the hypothesis that the eight haplotypes of malaria infected the children independently of each other. Thus, if it was known that a person was infected by a special type, this would not change the probability of being infected by any other special type. In subsequent analyses that followed, we used this result. The probability of being infected with the different types p_{XYZ} was estimated and is given in Table 1. If the different types were equally infectious, then the proportions of the eight types in the surrounding environment are given by f_{XYZ} estimates in the same table. The independence assumption might be checked in future analyses. If for example, the haplotype proportions in the infecting populations had been different at the two treatment locations, this would not have held for the mixed population, but as is seen in Table 1, the proportions (f_{XYZ}^* – columns) were quite similar. It might be interesting to study this further but larger populations and more locations will be needed.

At first recurrence of malaria, the proportions of some parasite types were smaller compared to baseline. In particular, those haplotypes with genes marked S at second and third positions decreased when treated with only SP. When the children were treated in combination with artesunate (ASP), the decrease was much smaller. This was seen from the α values in Table 5. One possible explanation could be that treatment with only SP did not kill all the parasites with resistant genes and that the surviving parasites were responsible for the reappearance of malaria. But treatment with SP in combination with artesunate, all were killed and all observed first recurrences depended on new infections. This is in agreement with the results in Kum et al. [12], where it was shown that there were more early recurrences when the children were treated with SP alone. There may be other explanations too like prophylactic effects.

We also noted that the parasites of type RRR with resistant genes at all three positions was the most common type. This indicates that there has been occurrences of resistant genes. The genes at the three positions were not independent. The proportion of parasites with three sensitive genes was high compared to the proportion with only two sensitive genes. This was verified by a formal test. This indicates that the proportion of R genes has not been the same for a long time. In that case one would expect equilibrium, that is, independence. This may be explained by the fact that there is an advantage in the accumulation of a second mutation and third one to structurally compensate the first one. The resulting more stable, three mutation carrying protein would be more readily selected under drug pressure.

Between 40 and 50% of the haplotypes in the infecting populations were the combination with only resistant genes RRR. Three other combinations RRS, RSR and SSS had more than 10% and the other four combinations were more rare. This may be due to an effect of earlier treatments (possibly with SP alone) that decreased all S genes simultaneously and increased the presence of the resistant genes at the three positions. The significant presence of SSS haplotype at baseline means (among other reasons) that this gene type had less contact with this treatment.

We also looked at the number of different haplotypes per child. It could be estimated as the sum of the values in the p_{XYZ}^* – column of Table 5 and corrected for having at least one infection. Since we could not reject independence between infections, we concluded that a child may be infected at two or more independent occasions with the same haplotype. A child in Kondé was on the average infected by 1.74 different parasite types, while in Uzini, a child was infected by 1.31 different types. The estimated number of times they were infected being 2.29 in Kondé and 1.46 times in Uzini, respectively. At the first recurrence of malaria, the number of haplotypes had decreased from 1.74 to about 1.20 types in Kondé. The average number of occasions a child was infected at this time point was 1.46. In Uzini, the number of haplotypes

had reduced to about 1.20, with 1.35 as the number of infection occasions, respectively. These occasions may have been before the trial and were not completely wiped out by the treatment.

In conclusion, our modelling procedure and the results obtained by applying these models to real clinical data contribute to knowledge and understanding of the dynamics of malaria parasite infections, drug efficacy and parasite genetics. These results are based on studies from two centres. It is impossible to tell whether these results can be generalised to all malaria types and all geographical locations. However, the methodology described in this article can be generalised.

Acknowledgements: The first author is thankful to the International Science Programme (ISP), Uppsala, Sweden for the financial support. We thank the anonymous referees for the comments and suggestions which improved the quality of this article. We equally thank the Division of Infectious Diseases, Karolinska University Hospital, Stockholm, for providing the data set.

Appendix

Table 6 Proportions of haplotypes at both centres

Genotype	KONDE		UZINI	
	Baseline	First recurrence	Baseline	First recurrence
	π_{ijk}^*	π_{ijk}^*	π_{ijk}^*	π_{ijk}^*
RRR	0.2535	0.6364	0.3232	0.4818
RRS	0.0634	0.0303	0.1970	0.1364
RSR	0.0986	0.1515	0.0656	0.0818
SRR	0.0000	0.0000	0.0000	0.0091
RSS	0.0000	0.0000	0.0101	0.0091
SRS	0.0000	0.0000	0.0051	0.0091
SSR	0.0000	0.0000	0.0000	0.0091
SSS	0.0211	0.0000	0.1212	0.0545
RRM	0.1338	0.0303	0.0404	0.0545
RMR	0.1056	0.0909	0.0808	0.0727
MRR	0.0141	0.0000	0.0000	0.0091
SSM	0.0000	0.0000	0.0000	0.0000
SMS	0.0070	0.0000	0.0051	0.0091
MSS	0.0000	0.0000	0.0000	0.0000
MRS	0.0070	0.0000	0.0000	0.0000
RMS	0.0070	0.0303	0.0000	0.0091
SRM	0.0000	0.0000	0.0000	0.0091
RSM	0.0070	0.0000	0.0000	0.0182
MSR	0.0000	0.0000	0.0051	0.0000
SMR	0.0000	0.0000	0.0000	0.0000
RMM	0.0915	0.0303	0.0451	0.0000
MRM	0.0000	0.0000	0.0152	0.0000
MMR	0.0141	0.0000	0.0000	0.0000
SMM	0.0000	0.0000	0.0000	0.0091
MSM	0.0070	0.0000	0.0101	0.0091
MMS	0.0282	0.0000	0.0303	0.0000
MMM	0.1408	0.0000	0.0451	0.0091
Healthy	–	109	–	88
log-mle	–316.5613	–40.10675	–411.0458	–206.7087

Table 7 Genotype classification and combinations

Possible classification		
Genetic element	Genotype	Possible combinations
R and S	–	RRR RRS RSR SRR RSS SRS SSR and SSS
1M	RRM	RRR + RRS
	RMR	RRR + RSR
	MRR	RRR + SRR
	SSM	SSS + SRS
	SMS	SSS + SRS
	MSS	SSS + RSS
	MRS	SRS + RRS
	RMS	RSS + RRS
	SRM	SRS + SRR
	RSM	RSS + RSR
	MSR	SSR + RSR
	SMR	SSR + SRR
2M	RMM	RRR + RSS, RSR + RRS, RSR + RRR + RSS, RSR + RRR + RRS, RSR + RRS + RSS, RSS + RRS + RRR, RSS + RRR + RRS + RSR
		RRR + SRS, SRR + RRS, SRR + SRS + RRR, RRR + RRS + SRS, SRR + SRS + RRS, SRR + RRS + RRR, RRR + SRS + SRR + RRS
		RRR + SSR, RSR + SRR, RSR + SSR + RRR, SRR + SSR + RRR, RSR + SRR + RRR, RSR + SRR + SSR, RRR + SSR + RSR + SRR
	SMM	SSS + SRR, SSR + SRS, SSS + SSR + SRR, SSS + SSR + SRS, SSR + SRS + SRR, SSS + SRR + SRS, SSS + SRR + SSR + SRS
		SSS + RSR, SSR + RSS, SSS + SSR + RSR, SSS + SSR + RSS, SSS + RSR + RSS, SSR + RSR + RSS, SSS + RSR + SSR + RSS
		SSS + RRS, SRS + RSS, SSS + SRS + RRS, SSS + SRS + RSS, SSS + RRS + RSS, SRS + RSS + RRS, SSS + RRS + SRS + RSS
	MMR	
	MSM	
	MMS	
3M	MMM	SSS + RRR, SSR + RRS, ..., RSS + RRS + SRR, ... 193 possibilities

References

1. Wigger L, Vogt JE, Roth V. Malaria haplotype frequency estimation. *Stat Med* 2013;32(21):3737–51.
2. Hastings IM, Smith TA. Malhaplofreq: a computer programme for estimating malaria haplotype frequencies from blood samples. *Malar J* 2008;7:130.
3. Hill WG, Babiker HA. Estimation of numbers of malaria clones in blood samples. *Proc R Soc Lond Ser B: Biol Sci* 1995;262:249–57.
4. Kessner D, Turner TL, Novembre J. Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. *Mol Biol Evol* 2013;30:1145–58.
5. Ross A, Koepfli C, Li X, Schoepfli S, Siba P, Mueller I, et al. Estimating the numbers of malaria infections in blood samples using high-resolution genotyping data. *PloS One* 2012;7:e42496.

6. Mårtensson A, Strömberg J, Sisowath C, Msellem M, Gil J, Montgomery S, et al. Efficacy of artesunate plus amodiaquine versus that of artemether-lumefantrine for the treatment of uncomplicated childhood plasmodium falciparum malaria in zanzibar, tanzania. *Clin Infect Dis* 2005;41:1079–86.
7. Kum CK. Bayesian analysis of two malaria treatments and probability of malaria parasite genotypes (PhLic thesis). Stockholm University, 2009.
8. Nelder J, Mead R. A simplex method for function minimization. *Computer J* 1965;7:308–13.
9. Burnham K, Anderson D. Model selection and multimodel inference: a practical information-theoretic approach. Springer Verlag: New York, 2002.
10. Cox D. Role of models in statistical analysis. *Stat Sci* 1990;5:169–74.
11. Pawitan Y. In all likelihood: statistical modelling and inference using likelihood. Oxford University Press: New York, 2001.
12. Kum CK, Thorburn D, Ghilagaber G, Gil P, Björkman A. A nonparametric bayesian approach to estimating malaria prophylactic effect after two treatments. *Int J Stat Med Res* 2013;2:76–87.