Evaluating accuracy of microsatellite markers for classification of recurrent infections during routine monitoring of anti-malarial drug efficacy: A computer modelling approach

Authors

Sam Jones, Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, United Kingdom; **Mateusz Plucinski,** Malaria Branch and U.S. President's Malaria Initiative, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America; **Katherine Kay**, Metrum Research Group , Tariffville, Connecticut, United States of America.; **Eva Maria Hodel**, Molecular & Clinical Pharmacology, University of Liverpool, Liverpool L69 3GF, United Kingdom; **Ian M. Hastings**, Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, United Kingdom.

Abstract

Anti-malarial drugs have long half-lives, so clinical trials to monitor their efficacy require long durations of follow-up to capture drug failure that may only become patent weeks after treatment. Reinfections often occur during follow-up so robust methods of distinguishing drug failures (recrudescence) from emerging new infections are needed to produce accurate failure rate estimates. "Molecular correction" aims to achieve this by comparing the genotypes between a patient's pre-treatment (initial) blood sample and any infection that occurs during follow-up, 'matching' genotypes indicating a drug failure. We use an *in-silico* approach to show that the widely used "match counting" method of molecular correction with microsatellite markers is likely to be highly unreliable and may lead to gross under- or over-estimates of true failure rates depending on the choice of matching criterion. A Bayesian algorithm for molecular correction has been previously developed and utilized for analysis of *in vivo* efficacy trials. We validated this algorithm using *in silico* data and showed it had high specificity and generated accurate failure rate estimates. This conclusion was robust for multiple drugs, different levels of drug failure rate, different levels of transmission intensity in the study sites, and microsatellite genetic diversity. The Bayesian algorithm was inherently unable to accurately identify low-density recrudescence that occurred in a small number of patients, but this did not appear to compromise its utility as a highly effective molecular correction method for analysing microsatellite genotypes. Strong consideration should be given to using Bayesian methodology for obtaining accurate failure rate estimates during routine monitoring trials of antimalarial efficacy that use microsatellite markers.

Background

Effective treatment of *P. falciparum* malaria infections is essential but is threatened by the spread of drug resistance to front-line antimalarial drugs, including artemisinin-based combination therapies (ACTs). Frequent monitoring of efficacy (1) is therefore necessary to confirm the effectiveness of current drugs, and to evaluate alternatives as they become available or necessary. The World Health Organization (WHO) recommends that endemic countries routinely re-test their currently used antimalarials at least every two years using standard patient-based in vivo trials. These are known as therapeutic efficacy study/studies (TES) (1); they generally only have one "arm" (i.e. the drug being evaluated). This terminology distinguishes TES from regulatory multi-arm trials used to evaluate proposed new regimens.

Current first-line antimalarials are ACTs which are composed of an artemisinin derivative and a partner drug; the artemisinin component is short-lasting and rapidly clears parasites, the partner drug has a longer half-life and is responsible for completely, but more slowly, clearing parasites. Ergo, if treatment fails due to resistance of parasites to the partner drugs, it may take several weeks for these drug failures to become patent. TES must therefore follow-up patients for several weeks post treatment to ensure failures are detected. The consequence of this requirement for long follow-up periods in these TES is that new infections not present at the time of treatment, termed "reinfections" (2) frequently occur during follow-up in moderate to high transmission areas. A patient presenting with detectable malaria parasites during follow-up (known as a recurrence / recurrent infection) may have a reinfection, which does not indicate that treatment failed to clear the patient's initial infection. Thus, recurrent infections trigger molecular testing that leads to them being classified as either: i) *Plasmodium falciparum* clones that infected the patient pre-treatment (initial clones) and were subsequently not cleared by treatment (termed recrudescence) or ii) reinfections that occurred during follow-up (3). In a methodology alternately called PCR correction or molecular correction, the genotype of the malaria infection at the time of treatment (the initial sample) is compared with the genotype of any recurrence during follow-up. The purpose of this comparison is to distinguish recrudescences from reinfections such that patients with reinfections can be excluded from subsequent analysis, thus producing a "corrected" drug efficacy estimate.

The original WHO and Medicines for Malaria Venture (MMV) consensus methodology was based on the use of length-polymorphic markers *msp-1*, *msp-2*, and *glurp*(3). An alternative system, microsatellite markers – segments of repeated genetic motifs - has been explored (4-6), a proposed advantage being the lack of immune selection on ostensibly neutral microsatellite markers (7). In this methodology, researchers genotype microsatellite loci in both initial and recurrent infections and count the number of matching loci in each patient i.e. the number of loci at which at least a single allele is shared between the initial and recurrent infection. They then define a certain number of matches to be indicative of recrudescence. In addition to their use in TES, microsatellites have also been commonly used to assess treatment failure in returning travellers in non-endemic areas (8-10).

There are two inherent sources of bias in running TES, independent of the genotyping method:

a) A patient who fails to clear their initial infections may have a reinfection that becomes patent before the recrudescent clone reaches a detectable level; ethically, that patient must be treated and so is removed (or "censored") from the study before the recrudescence can be observed.

b) A patient who fails to clear their initial infection may have that infection persist at a low-lying level, below the limit of detection of light microscopy (assumed, see later, to be 10^8 total parasite count in the patient), such that parasites are never detected during follow-up; the frequency of this event is influenced by the duration of follow-up in the trial i.e. the longer the follow-up, the less likely it is to occur.

Classification of recurrences into reinfections and recrudescences also introduces potential bias into estimates of the true failure rate (11-14). Genotyping of blood samples is imperfect and suffers from three key limitations. Firstly, patients are often infected by two or more malaria clones in high transmission areas and the lower density "minority clones" contribute genetic signals that are hard to detect in the amplification process, meaning their low frequency alleles may fail to be detected in either blood sample (4). Secondly, there are inherent error rates when measuring the genotype of parasite, for example through imperfect determination of fragment length or sequencing error. Finally, there is the non-technical limitation that reinfections can, by chance, share alleles with clones of the initial infection – this is more pervasive in areas of lower genetic diversity (6). These three factors combine to generate several additional several factors that need to be considered in the analysis of malaria drug trials:

c) Recrudescent infections can be misclassified as reinfection if alleles of the recrudescent clone were not detected when genotyping the initial infection.

d) Recrudescent infections can be misclassified as reinfection if a recrudescent clone has a sufficient number of base pair read errors (i.e., at multiple markers) such that it appears to be a reinfection.

e) A reinfection can be misclassified as recrudescent if it shares (by chance, or due to base pair read error) a genetic signal(s) with those clones present at time of treatment.

Typically, microsatellite data are analysed by applying a mathematically simple match counting algorithm which uses an arbitrary threshold for the number of loci that have common alleles between the initial and day of failure samples. In these algorithms, if the two samples have matching alleles at, or above, the threshold number of loci, they are classified as recrudescence, and otherwise, reinfections. Typically, classification of an infection as a recrudescence requires a match at most, if not all, sampled loci (4, 15, 16). This kind of counting algorithm only deals with the unprocessed, "raw" genetic data and makes no allowance for errors due to factors c to e described above resulting in increased risk of misclassification, although more advanced statistical algorithms have been proposed to adjust for these potential biases (6).

A recent publication (17) presented a statistical method based on Bayesian probability to analyse microsatellite data to calculate drug failure rates. This method generates the posterior probability that a recurrent infection is a recrudescence and has subsequently been used to analyse TES data (18-20). The biases listed above mean that a simple method of counting matching microsatellites between samples may never be able to reliably classify a patient as reinfection or recrudescence. Bayesian analyses constitute a better, more flexible approach capable of dealing with these uncertainties and the advantages of a Bayesian approach are explained elsewhere (17).

We utilized a computer modelling approach to simulate therapeutic outcome following antimalarial therapy in anti-malarial trials. In these simulated data-sets, the parasitaemia of each patient's clones is calculated at every time-step, thus the true status (reinfection or recrudescence) of all recurrent infections is known. Using the simulated data, we then evaluated the ability of the Bayesian algorithm to correctly distinguish reinfections from recrudescences. This allowed us to quantify the accuracy of this method, which has not been possible *in vivo*; due to imperfect molecular correction techniques, the true failure rate of the population cannot be known. We also compared the performance of the Bayesian algorithm to a threshold-based match counting algorithm, and investigated whether the advantages of a Bayesian methodology are realised in the analysis of data from anti-malarial TES and whether this approach is truly as "robust" as postulated in the original paper (17).

This study had three main objectives: Firstly, to evaluate the accuracy of failure rate estimates generated using microsatellite data in conjunction with a match counting algorithm (as is currently typical). Secondly, to assess the advantages of Bayesian analysis methodology, both in its ability to recover the true failure rate and the diagnostic ability to distinguish recrudescence from reinfections. Thirdly, to check whether the methodologies based on microsatellite loci are robust across drugs with different post-treatment prophylactic profiles (i.e., partner drugs with varying half-lives) which determine when reinfections start to occur, across different transmission intensities (which determine rates of reinfection in TES) and in regions with differing levels of genetic diversity at microsatellite loci.

Methods

Study Design

We used existing pharmacokinetic / pharmacodynamic models (PK/PD) (21-23) to simulate parasite intra-host dynamics following treatment in 10,000 patients. We simulated whether original clones were cleared or survived drug treatment. If they survived we then noted, whether the recrudescent clone(s) became patent during follow-up, and if/when reinfections occurred and became patent. We allowed clones present at time of treatment to have different numbers (densities) and assigned microsatellite alleles to each clone in the infection. This allowed us to simulate the genetic information that would occur during routine follow-up in these simulated patients that reflect the inherent problems in the follow-up and genotyping processes (i.e., inability to detect low density clones, and genotyping errors as described above).

We ran 12 different scenarios, varying the drug, the failure rate, and the level of transmission intensity, The latter factor is quantified as the force of infection (FO]) which is the frequency at which reinfections emerge per person per year. Transmission intensity also affects the initial number of clones in each patient at time of treatment (commonly known as the multiplicity of infection (MOI) or, equivalently, complexity of infection), and the level of genetic diversity in the population allele structure (details below). For each scenario, we used the Bayesian algorithm and the match-counting algorithm to generate estimates of the failure rate. We then compared the estimate of the failure rate with the true failure rate to assess the accuracy of each algorithm. It is important to note that our methodology has two distinct steps: First, the mPK/PD model simulates parasite dynamics post-treatment, and we used a series of heuristics to calculate which alleles would be observable at any given time-steps. This provided data-sets that are akin to those obtained *in vivo*, where each patient's infection is described by observable alleles in the initial sample and observable alleles of any recurrent sample. Secondly, we applied the match-counting and Bayesian

algorithms to this data to obtain failure rate estimates. The simulated data-set could then be used to analyse algorithm performance against the true failure rate (known from the simulation).

Computational methodology

All modelling and subsequent analysis was conducted using the statistical programming language R (version 3.5.1) (24). Figures were produced using base R graphics, and the *ggplot2* package (25). For hardware details and programming considerations, see **[Supplementary Information, SI]**.

Trial Scenarios

Twelve TES scenarios were simulated. The main body of this manuscript presents results obtained from simulations of artemether-lumefantrine (AR-LF) therapy, with results for the case of artesunate-mefloquine (AS-MQ) presented in the **[SI]**. The purpose of simulating two distinct treatments reflects the different post-treatment prophylactic duration of the drugs – AS-MQ persists at killing concentrations for longer than AR-LF. We primarily wanted to analyse the use of microsatellite markers for AR-LF treatment for which a Bayesian approach has been previously been applied (17), but we also wanted to test if results were consistent for a drug with a longer post-treatment prophylactic period. For each drug, we simulated non-failing drugs with low failure rates (1-2%) and failing drugs with higher failure rate (~10%). True failure rate of the drug is determined by the half maximal inhibitory concentration (IC50) of the drug in the parasite population; the IC50 of each clone is drawn from a distribution of values (Table S1 of **[SI]**). Note that we arbitrarily changed the mean IC50 value of partner drugs within the model to obtain different levels of treatment failure. We do not imply the values of IC50 here are representative of any particular field scenario, but rather use them to investigate the accuracy of techniques to analyse clinical trial data in a simulation environment. In our simulations, true failure rate changes with MOI (a higher MOI means that there are more clones within a patient that are potentially able to survive treatment, and so true failure rate increases), so we altered mean IC50 between scenarios for failing drugs to keep true failure rates within a percentage of each other between scenarios. Each drug calibration (i.e., non-failing AR-LF, failing AR-LF, non-failing AS-MQ and failing AS-MQ) were run in low, medium and high transmission settings. These scenarios incorporated varying distributions of multiplicity of infection (MOI) at time of treatment, different frequency distributions of microsatellite alleles (obtained from Angola and based on transmission level, see parts 2.3 and 2.4 of **[SI]**), transmission intensity (quantified by FOI; see part 2.2 of **[SI]**) The calibration of each scenario in terms of MOI, allele frequency, FOI and mean IC50/True failure rate is presented in [**SI]**. Each scenario simulated 10,000 patients for a 28-day follow-up period for AR-LF and as 42-day follow-up period for AS-MQ.

The data we used for distributions of microsatellite markers came from three sentinel sites in Angola (17, 18), which represent areas with moderate to high transmission and thus relatively high diversity (Part 2.3 of [**SI**]). The risk of misclassifying a reinfection as a recrudescence is, intuitively, higher in areas of lower genetic diversity (potential error (e) described in the background), so we artificially generated an additional distribution of marker allele frequency with very low genetic diversity by modifying the allele distributions from the low diversity area (as described in part 2.4 of **[SI]**)to investigate the accuracy of failure rate estimates under this condition.

PK/PD model specifications and output

We utilized a computer-based mechanistic PK/PD model of drug treatment of uncomplicated *P. falciparum* with either AR-LF or AS-MQ based on previous models (21-23). The methodology used the drug concentration profile in each patient to calculate the change in parasite counts (parasitaemia) of each malaria clone over time following drug treatment; this produced quantitative estimates of parasite dynamics in a patient following treatment (Figure S1, [**SI]**). The drug concentration over time in the patient population for each partner drug is shown in (Figures S2 and S3 **[SI])**. An alternative approach is to generate parasite dynamics by arbitrarily deciding on a day of recurrence for a patient, then assigning the recurrence as containing recrudescences and/or re-infections (e.g.(26)). It would then be straightforward to draw the parasite numbers in each clone from a uniform distribution but a PK/PD model was chosen for the ability to easily test different levels of drug failure, for increased realism, and to allow future users to easily re-calibrate this methodology with parameters of their choice. Pharmacokinetic parameters varied between patients and pharmacodynamic parameters varied between clones by drawing them from distributions of PK and PD parameters (Table S1 [**SI])** for full parameter lists, and additional considerations.

The number of initial clones in each patient was drawn from a distribution of multiplicity of infection (MOI) that depends on local transmission intensity (MOI ranges between 1 and 5; see Figure S4 of **SI**]; the starting parasitaemia of each clone was drawn from a log-uniform distribution between 10^{10} and 10^{11}. We describe parasitaemia in terms of parasite counts, rather than parasite densities. We do this because the models track changes in parasite counts over time and we do not parameterize patients in such a way that would allow us to easily convert counts to parasite densities (i.e., patients do not have parameters for blood volume, white blood cell (WBC) count or red blood cell count, etc.), nor would including these parameters aid the mechanistic simulation of the model or improve the accuracy of the results. For reference, assuming a patient with 4.5L of blood and a WBC count of 8,000/µl of blood, parasitaemia of 10^{10} and 10^{11} would correspond to densities of 2,222 parasites/µl of blood and 22,222 parasites/µl of blood respectively, per WHO counting procedure (27). Previous modelling approaches used 10^{12} parasites as the upper limit of parasitaemia; this level of parasitaemia is likely to be lethal or at least exceed the maximum parasite density exclusion criteria in a clinical trial (typically 100,000 parasites /µl); hence we used 10^{11} as the upper limit for any single clone at the time of treatment. The number of reinfections that occurred during follow-up was determined by the parameter FOI which we used as our measure of transmission intensity (Section 2.2 of [**SI**]). The days on which reinfections occurred was drawn from a Poisson distribution whose mean was the FOI. Reinfections were assumed to emerge from the liver at a count of 10^5 parasites (28, 29) [**SI**]. PK parameters were varied between patients and PD parameters were varied between malaria clones such that each patient and clone responded differently to treatment (see [**SI**], Table S1 for table of PK/PD parameter means and associated coefficients of variation [CV]). The growth rate of each clone was assumed to be identical for every clone and set to 1.15/day as in previous modelling work (23, 30); this is equivalent to a parasite multiplications rate of 10 per 48 hour cycle. The simulation assumed that if the total parasitaemia (i.e. the sum of parasitaemia of all clones) in a patient at any time, reached 10^{12}, then density-dependent effects, such as fever, acted to control and stabilise the parasitaemia, effectively setting the growth rate of every clone in that patient to 0. Aside from this density-dependent effect, we did not attempt to model patient acquired immunity as accurately modelling this acquisition is notoriously difficult. It is likely to affect recrudescent and re-infecting clones equally such that we would not expect it to alter how recurrent infections are classified. We did not model parasite sequestration (see (31) for justification). The output of this PK/PD model was, for each patient, the exact number of parasites of each malaria clone (be that clone an initial infection or a reinfection) at each time-step of the model (days); see figure S1 **[SI**] for an example. A patient in a real TES would be removed from the trial and re-treated

when a recurrence occurred, but no such ethical imperative exists *in silico* so we tracked the patients the full length of follow-up, with the advantage that we could determine if any initial clones were still present on the final day of follow-up, even though, *in vivo*, that patient may have been removed from the trial (right-censored) earlier due to a recurrence caused by reinfection.

Modelling microsatellite genotyping and detectability of alleles

Genotypes were assigned to every clone (both initial and reinfections) at seven microsatellite markers: *313*, *383*, *TA1, polya, PfPK2, 2490* and *TA109*; alleles at each marker were defined by their length (base pairs), see details in part 2.3 of **[SI]**.

The genotype of the initial malaria infection of each patient was taken on the day of treatment. This genotype signal is a composite of all the clone(s) present in the initial infection and is determined by the technical accuracy and sensitivity of genotyping (points (b) and (c) in the Background and see later). Each patient was then checked for recurrent parasites on days of follow-up in a typical clinical trial schedule i.e. day 3, 7, 14, 21 and 28 for AR-LF and additional days 35 and 42 for AS-MQ.

On all days of follow-up except day 3, a recurrence was identified if the sum parasitaemia of all clones in a patient exceeded 10^8 which we assumed was the minimum parasitaemia at which detection by light microscopy was possible (32). This corresponds to a parasite density of roughly 22 parasites/µl of blood assuming a patient with 4.5L of blood and 8,000 WBC/µl. If total parasitaemia was less than 10^8 then recurrent parasites would not be observed by microscopy (and thus, the patient would not be genotyped on that day). On day 3, if total parasitaemia exceeded 10^8 but was <25% of the total parasitaemia on the initial sample, the patient continued in the trial; if parasites were present at >25% of initial parasitaemia, that patient was classed as an early treatment failure, per WHO procedure (1);. Genotyping of initial and recurrent samples was then simulated using the following 3-stage protocol:

Firstly, we included a "sampling" limit: A finite volume of blood is available for genotyping. A parasite clone would not be detected if its density were so low that no parasites were included in the blood sample analysed. Thus, the density and volume of the processed blood sample defined the limit of detection. We assumed this limit to be 10^8 (i.e., no clone present in less than 10^8 parasites would be detected); see part 3 of **[SI]** for calculation and justification.

Secondly, the "majority" allele for each microsatellite is the allele with the highest parasitaemia (if multiple clones share alleles at a marker, the allelic signal for that marker is the sum of parasitaemia of the clones). We assumed that for an allele to be detected, the parasitaemia of that allele must be ≥25% of the parasitaemia of the majority allele; this reflects the sensitivity of microsatellite genotyping to infer low-frequency alleles.

Finally, we included the chance that the length of each microsatellite may be mis-read due to genotyping errors such as stutter bands (7) . The chance of an error of +/- length *x* was assumed to be described by the geometric distribution $0.8 * (0.2)^x$, described in (17).

The output of these simulations was, for each patient, the microsatellite alleles (quantified by their length in base-pairs for each loci) at each of the seven loci, observed in the initial sample, and at any recurrent infection in that patient. A small example (100 patients) is shown in **[Supplementary file 1]** This is exactly the data recorded in standard TESs (and is the input used for the Bayesian algorithm *in vivo* as in (17, 18)) so this data formed the basis for our PCR-correction and failure rate estimates.

296 ## *Terminology of results*

297 Our terms “Recurrent infection/Recurrence”, “Recrudescence” and “Reinfection” are consistent with
298 the WHO terminology (2).

299 We frequently use the additional term “true failure” to describe the failure rate that we know
300 occurred during our simulations (and which is unknown in a real TES). We determined whether each
301 patient was a “true failure” based on parasitaemia: A patient was a true failure if, on the final day of
302 follow-up (day 28 for AR-LF, day 42 for AS-MQ), they still harboured any parasites from any initial
303 clone. The true failure rate is the frequency of these patients across the entire population. Our
304 model tracked patients over the full length of follow-up, thus our “true failure” classification
305 captured patients who would, in a real trial, have been removed earlier in the trial with a recurrent
306 infection classified as a reinfection (and whose recrudescent clones would not then be observed).

307 A key advantage of our *in silico* approach is its ability to interrogate the Bayesian algorithm; i.e.,
308 investigate diagnostic ability and determine in which circumstances it would misclassify recurrences.

309 For these analyses, we separated true failures into ‘high’ and ‘low’ density recrudescence. The
310 performance of PCR correction is likely to depend on its ability to detect genetic signals from low-
311 density clones. The detection limit for low-level genetic signals in our simulation was 25% (to reflect
312 current genotyping sensitivities, described above) so its is useful to compare the methodologies
313 when patients have high-density recrudesce (recrudescing clones are present at >25% in both initial
314 and recurrent samples) and low-density clones. Technically, a high density recrudescence was
315 defined as occurring when three conditions were met: (i) if there is a mixed infection of new and
316 recrudescent clones on the day of recurrence, recrudescent clones must be >25% of the total
317 infection (more specifically, the sum parasitaemia of all recrudescent clones on the day of
318 recurrence must be >25% the sum parasitaemia of *all* clones on the day of recurrence)) and (ii),
319 Clones that recrudesce must constitute at least 25% of the initial infection (more specifically, the
320 sum parasitaemia of all recrudescent clones on the day of recurrence must have been >25% of the
321 total parasitaemia of all clones in the initial sample).) (iii) the total number of parasites in
322 recrudescing clones on the day of recurrence must be $\geq 10^8$ (to be consistent with the sampling limit
323 defined above). If any one of these conditions is not met then the failure is defined as “low density”.
324 In this manner, we determined the true classification of each recurrence as a reinfection, high
325 density recrudescence or low density recrudescence.

326 ## *Match counting algorithm*

327 A match counting algorithm compared the number of microsatellite loci that have at least a single
328 allele shared between the initial and recurrent sample (termed a “matching” loci). Typically, use of
329 microsatellite markers *in* vivo requires a high number of matching loci to classify an infection as
330 recrudescent (either all loci, or permitting a single locus not to match, i.e.: (4, 15, 16)). Herein, with
331 the 7 loci modelled, we vary the threshold number of matching loci required to classify a
332 recrudescence to determine the impact of this choice of threshold on failure rate estimates. This is a
333 counting algorithm where a recurrent infection is defined as a recrudescence when the number of
334 matching loci is greater than or equal to a specified threshold. Six threshold values were analysed for
335 this method: 2, 3, 4, 5, 6 and 7 matching loci (e.g. if a recurrent infection had 3 matching loci with
336 the initial infection, that recurrence was classified as a recrudescence with a threshold of 2 or 3 loci,
337 but as a reinfection with the other thresholds.

338 ## *Bayesian analysis method*

We used the Bayesian analysis method described in (17) to interpret our simulated results and obtain posterior probabilities of recrudescence for each patient. In brief, the Bayesian algorithm uses a Markov chain Monte Carlo approach to sample from the posterior probability of recrudescence for each sample, with the ratio of likelihoods of a reinfection versus a recrudescence derived from the frequencies of the observed alleles. The algorithm jointly estimates several key parameters, such as the genotyping error rate, and accounts for missing data by sampling hidden alleles. The data input into the Bayesian algorithm in our simulations is the same as occurs in analysis of *in vivo* trial data, i.e., the microsatellite profile of initial and recurrent infections in each patient as shown in **[Supplementary file 1].**

The Bayesian analysis was then used to define a recurrence as being a recrudescence when posterior probability of recrudescence in that patient exceeded a value *p*, where *p* lies between 0 and 1.

Note that the Bayesian algorithm is applied to our simulated data-sets in the same way it is applied to *in vivo* data (described in (17)). Crucially, this means that the priors for all parameters are uninformative – we are not calculating any given parameter in the mPK/PD framework and then using that parameter as a prior for the Bayesian algorithm (which would clearly invalidate results). Note, though, that posterior estimates from (17) are used to inform the chance of allele length being mis-read in the mPK/PD model, described above.

Assessment of algorithm accuracy

Both the match-counting algorithm and Bayesian analysis classified a recurrent infection as either reinfection or recrudescence depending on the choice of threshold (for the match counting algorithm) or posterior probability *p* (for the Bayesian analysis). These classifications were then used to generate failure rate estimates for the simulated TES using survival analysis (the WHO-recommended method (1)) with the *R* packages *survival* (33) and *survminer* (34) . The failure estimates for both methods were then compared with the true failure rate to assess their accuracy.

The distribution of the posterior probability of recrudescence calculated using the Bayesian algorithm was plotted for each scenario, with recurrences stratified into their true status: low-density recrudescence, high-density recrudescence or reinfection. Receiver operator characteristic (ROC) curves were constructed using the posterior probability at which an infection would be classified as a recrudescence (from 0 to 1). The area under the ROC curve (AUC) was used to quantify the diagnostic ability of the method (35), with an AUC of >0.8 considered to be a "good" test and an AUC of >0.9 considered to be an "excellent" test.

We evaluated the ability of the Bayesian algorithm to detect low-density recrudescence, by calculating the posterior probability of recrudescence estimated by the Bayesian algorithm for each recurrent infection and categorizing each infection as reinfection, low-density recrudescence or high-density recrudescence as described above.

Results

Results were generated for AR-LF and AS-MQ under the assumption of a failing and non-failing drug for three scenarios of transmission intensity (methods). Here we focus on the results for AR-LF while the results for AS-MQ are fully described in **[SI]**.

Failure rate estimates and comparison to true failure rate.

The match counting algorithm was sensitive to transmission intensity; no threshold value of matching loci at which a recurrence was classified as recrudescent was able to accurately estimate true failure rate across all transmission scenarios for either failing (Figure 1) or non-failing (Figure 2) AR-LF. Failure rate estimates declined as the threshold increased. Failure rate estimates increased as transmission increased, presumably due to the greater number of reinfections, some of which will be misclassified as recrudescences; this effect was greater at low thresholds when the probability of such misclassification was greater. A threshold of 4 matching loci produced estimates close to the true failure rate for all non-failing AR-LF scenarios. For failing AR-LF scenarios, a threshold of 3 matching loci produced the closest estimate to true failure in the low transmission scenario, and a threshold of 4 matching loci produced the closest estimate in the high transmission scenario, with the medium transmission scenario intermediate between the two. However, using a threshold of 3 matching loci in a high transmission scenario over-estimated failure rate (estimated failure rate of 0.18 compared to a true failure rate of 0.1). A threshold of 4 matching loci gave an estimate of 0.08 relative to a 0.0997 true failure rate for the failing, medium transmission scenario and an estimate of 0.077 relative to a true failure rate of 0.0965 for the failing, low transmission scenario. A threshold of 7 matching loci resulted in extremely large under-estimates of failure rates for failing AR-LF: 0.005 relative to true failure rate of 0.0965 in the low transmission scenario, 0.008 relative to true failure of 0.0997 in the medium transmission scenario and 0.006 relative to true failure rate of 0.1 in the high transmission scenario.

In contrast to the match-counting method, the Bayesian algorithm recovered true failure rate to a high degree of accuracy across all transmission settings and for both calibrations of true drug failure rate (Figure 1 and Figure 2). Values of the posterior probability of recrudescence, p, used to distinguish recrudescence from reinfection between 0.1 and 0.9 produced good, consistent failure rates estimates with only a slight decline as p increased; using $p = 1$ to classify a recrudescence resulted in a substantial decrease in failure rate estimates. For all non-failing and failing drug scenarios, treating all infections with $p \geq 0.1$ as recrudescence generated a failure rate estimate within 0.01 of the true failure rate.

Receiver Operator Characteristic (ROC) curves for the Bayesian algorithm.

The general trend was that the AUC of the ROC curve decreased as transmission intensity increased (Figure 3), with values of 0.872 and 0.835 in the failing and non-failing high transmission scenarios respectively – these correspond to a "good" diagnostic test. AUC was higher for any given transmission scenario in failing AR-LF than non-failing AR-LF. When the ROC curve was calculated for only high-density recrudescence AUC increased to ≥0.968 in all scenarios – an "excellent" diagnostic test.

Distribution of posterior probability of recrudescence

Figure 4 shows the distribution of the posterior probabilities of recrudescence for all recurrences, stratified according to the true classification of their recurrence: Reinfection, low-density recrudescence, or high-density recrudescence. The distributions were nearly binary in every scenario: Nearly all posterior probabilities in the patient population were <0.1 or≥0.9. Some trends here were intuitive (note different scales on the Y axes) : i.e., larger number of reinfections occurred as transmission intensity increased and larger number of recrudescences occurred in scenarios in which failing drugs were administered. The small number of patients whose infections had estimated probabilities of recrudescence between (but not including) 0.1 and 0.9 was reflected in the minor changes in failure rate estimates as p changed in Figure 1 and Figure 2.

Most patients whose recurrence had p <0.1 were reinfections. Given that ≥0.1 was the choice of p that produces the most accurate failure rate estimate (Figure 1 and Figure 2), the cause of the (slight) under-estimate of failure rate was due to the proportion of patients with infections at p <0.1 who had, in reality, recrudescent infections. In simulations of failing drugs, at all transmission intensities, ~5% of recurrent infections that had p <0.1 were truly recrudescent infections . For simulations of non-failing drugs, at all transmission intensities , ~2.5% of recurrent infections that had p <0.1were truly recrudescent infections. Notably most of these were low density recrudescence; only 0.03%-0.05% of recurrent infections that had p <0.1 were high-density recrudescences in simulations of failing drugs, and 0.02%-0.06% of recurrent infections that had p <0.1 were high-density recrudescences in simulations of non-failing drugs . There were a small number of recurrent infections with p ≥0.1 which were truly reinfections but in all scenarios this number was small relative to the number of recurrent infections that had p <0.1 and were truly recrudescent.

Consequently, the under-estimation of failure rate that occurs due to truly recrudescent infections having p <0.1 was greater than the over-estimation due to reinfections having p ≥0.1; thus these reinfections with p ≥0.1 were not leading to an over-estimation of failure rate.

Figure 1, Figure 2 and Figure 4 show that over-estimation of failure rate due to misclassification of reinfection as recrudescence did not significantly affect the Bayesian algorithm due to its high specificity; nearly all reinfections have a posterior probability of recrudescence of <0.1. A slight-under-estimate of failure rate occurred with all values of p ≥0.1-≥0.9 inclusive to classify a recrudescence, due to the algorithm assigning posterior probabilities of <0.1 to a small proportion of infections with low density recrudescence.

Determinants of posterior probability of recrudescence

Figure 5 is a contour plot showing the estimated posterior probabilities of recrudescence estimated by the Bayesian algorithm as a function of the densities of the recrudescent clone(s) in the recurrent and initial sample. There was a clear trend of the posterior probability of recrudescence increasing as both densities increase, reinforcing the result illustrated in Figure 4: the density of recrudescent clones was an important determinant of the posterior probability of recrudescence returned for a given patient. Errors in classification were due almost entirely to the finite sensitivity of genotyping causing some low-density clones to be missed during genotyping.

Analysis of Artesunate-Mefloquine

We simulated and analysed AS-MQ in the same manner as for AR-LF. Full results are shown in **[SI]**. Results were very consistent with those of AR-LF: The match counting algorithm for classifying recurrences as reinfection or recrudescence could not consistently provide accurate failure rate estimates across a variety of scenarios and often resulted in extreme over or under-estimates of true failure rate, depending on the choice of threshold. The Bayesian analysis method generated failure rate estimates to a high degree of accuracy across all scenarios, although there was an under-estimate of 1.6 percentage units in the high transmission, failing drug scenario. As with AR-LF, using $p \geq 0.1$ to classify an infection as a recrudescence provided the most accurate failure rate estimate for AS-MQ in every scenario.

Very low genetic diversity scenario

As expected, in the very low genetic diversity scenarios, failure rate estimates increased due to misclassification of reinfection as recrudescence (factor I) identified in the Background) **[SI]**. The match counting algorithm was unable to recover accurate failure rate estimates in any scenario. However, using a high threshold of matching loci to classify a recrudescence (6 or 7) did not lead to over-estimates of failure rate, even in a high transmission setting, under conditions of very low genetic diversity. Importantly, the Bayesian method recovered accurate failure rate estimates in low genetic diversity scenarios when using $p \geq 0.1$ to classify a recrudescence **[SI]**.

Patients with sub-patent, undetectable parasitaemia during follow-up

We calculated the number of patients who had undetectable, sub-patent parasitaemia on the final day of follow-up (i.e. a total $<10^8$ parasites, either reinfection or recrudescence), and the proportion of these patients who were harbouring sub-patent recrudescent infections. These results are shown in full in Table S3 of [**SI**]. The results we present above are based on analysis of patients with patent recurrent infections (i.e. those who have detectable parasites during follow-up). In our model, it is possible for a patient to be a true failure (i.e., fail to clear their initial parasite clones), but never have detectable levels of parasites (either recrudescent clones *or* reinfections), during follow-up (methods). If the number of these patients were large, it would induce bias in our results. However, the proportion total patients who were true failures but had no recurrent infection was extremely low (between 0 and 0.001 across all scenarios for non-failing and failing AR-LF), so we can assume duration of follow-up, at least in our simulations, was sufficiently long to capture nearly all failures and hence safely draw conclusions about the entire study population.

Discussion

Our *in silico* experiment showed that the Bayesian algorithm generated extremely accurate estimates of true failure rate across different transmission intensity and drug failure rate scenarios. In contrast, the match counting algorithm showed high potential for misclassification bias, with no single threshold able to consistently estimate the true failure rate.

Our results highlight the important role that computer modelling approaches can play in evaluating the performance of genotyping-based classification algorithms. This kind of approach is essential for this evaluation because, unlike real field data, we know the true failure rate of drugs *in silico* so can readily identify the most accurate and/or robust method of analysis. In contrast, analysis of field data demonstrates that failure rate estimates vary depending on choice of methodology (e.g., between criteria used to define recurrence as reinfection or recrudescence, i.e. (11)) but, since the real failure rate in a clinical trial is unknown, it is not possible to demonstrate which method is most accurate or robust. A further advantage of simulated datasets is that we can observe the conditions under which a method fails to return a correct classification (for example, Figure 4). We are confident in our conclusions for several reasons.

Our first main conclusion is that despite its wide use, match counting of microsatellites for distinguishing recrudescence from reinfection is not a robust approach because the estimated drug failure rate is highly dependent on the threshold used to define a recrudescence. By definition the same clone of malaria will have the same genotype between the initial and recurrent sample. However, the *observed* genotype (described by the microsatellite alleles) may differ due to issues inherent in the genotyping method (failure to detect minority alleles or errors in measuring base-pair length of alleles) – accounting for this difference is the purpose of including a degree of

flexibility in the molecular correction process i.e., varying thresholds. Use of microsatellites to correct trials *in* vivo has, up to this day, generally relied upon a simplified analysis method such as the match counting algorithm described here. Hwang et al. (16) used 8 markers and defined a match at 7 or more loci to be a recrudescence. Greenhouse et al. (4) investigated 6 markers, and subsequently used 4 to analyse samples, with a match at every locus being required to classify a recurrence as a recrudescence. Mwangi et al. (15) used 5 loci and considered a match at 5 to be a recrudescence, 0 to be a reinfection, and intermediary values to be mixed infections.

The high thresholds generally used to classify a recurrence as a recrudescence (either most, or all, of the available loci must match to define a recrudescence) likely results in substantial under-estimate of failure rate. For the *in silico* failing AR-LF results presented here, failure rate estimates with a threshold of 2 ranged between 15% in a low transmission scenario to 50% in a high transmission scenario, relative to true failure rates of ~10% (Figure 1). However, a threshold of 7 provided estimates that ranged between 0.5% and 0.6% relative to true failure rates of ~10% . For non-failing AR-LF (Figure 2) failure rate estimates with a threshold of 2 ranged from 7% in a low transmission scenario to 24% in a high transmission scenario, relative to true failure rates of ~2%. In other words, the potential bias induced by choice of a break-point for the match counting algorithm could result in either rejecting an efficacious drug or continuing to use a failing drug and this is further complicated by the sensitivity of the break-point to transmission intensity (Figure 1 and Figure 2); the same issues are present in using the match counting algorithm for AS-MQ **[SI]**. This is perhaps not surprising: In the context of genetic markers for classification of recurrent infections in TES, microsatellites are very similar to the marker *glurp* used in the WHO/MMV method i.e. are defined only as length polymorphism with no allelic families. This has led some commentators to suggest glurp is so unreliable that it should simply be omitted from the WHO/MMV method or used only to resolve disparate *msp-1* and *msp-2* results (11).

The results presented here strongly suggest that stringent thresholds (i.e., requiring all or most loci to have matching alleles) will under -estimate failure rate (and over-estimate efficacy). With the seven microsatellites used in these simulations, failure rate estimates produced by the match counting algorithm varied with both the choice of threshold and the transmission intensity but in all scenarios a threshold of 5 matching loci under-estimated failure rate; either 3 or 4 produced the closest estimate (Figure 1 and Figure 2). A threshold of 2 would lead to large over-estimates of failure rate. The reason that stringent thresholds under-estimated failure rate is because low-density recrudescence can be overlooked in patients who have a polyclonal initial or recurrent infection. Note that the threshold producing the most accurate estimate increased from 3 to 4 as transmission increased from low to high – this is because in higher transmission areas there was a greater impact of reinfections incorrectly classified as recrudescence due to sharing alleles by chance. However, this will be dependent on the genetic diversity of the markers used.

When a match counting algorithm for interpreting microsatellite data is used, we strongly suggest that failure rates obtained with multiple thresholds points are reported, (for example Plucinski et al. reported failure rate estimates based on thresholds of matching at all loci and matching at all except a single loci; their table 2). This reflects the difficulty (in our opinions, the impossibility) of identifying a robust threshold (our figures 1 and 2) *a priori*. Additionally, we suggest that stringent thresholds (requiring all or a very high proportion of loci to be matching) are generally avoided. Inaccuracies of failure rate estimates using the match counting algorithm were a concern for failing drugs; a threshold of 4 was a reasonable approach in our non-failing drug scenarios as most recurrences were likely to be reinfection and 4 appeared to be a sufficient threshold to prevent over-estimation of failure rates due to misclassifying reinfections as recrudescence. Consequently, a

feasible approach for using microsatellites in TES would be to use the match counting algorithm initially, assess the failure rate estimates produced with a range of thresholds and pass any result that indicates a drug failure rate of higher than 5% through a Bayesian algorithm for re-analysis. We note that in our results, the estimates produced by each threshold are sensitive to transmission intensity, but even in a high transmission intensity, a threshold of 4 would not mistakenly indicate that a failing drug was non-failing (Figure 1).

Our second main conclusion is that application of the Bayesian algorithm produces relatively accurate and stable estimates of failure rate in all transmission scenarios for both failing and non-failing with use of a posterior probability *p* of 0.1. This result is consistent for analysis of AR-LF, AS-MQ and even for AR-LF in a very low genetic diversity setting, where a *p* of 0.1 is effective due to the high specificity of the Bayesian algorithm, i.e., misclassification of reinfection as recrudescence is extremely infrequent (Figure 4 and **[SI]**). However, note that in the very low diversity setting, failure rate estimates increased as transmission intensity increased, and in areas of higher transmission than we simulated here there may be a risk to accurate classification with this method; though this pre-supposes that low genetic diversity (characteristic of low-transmission settings) could occur within an area of high transmission.

The type of PK/PD modelling that we used to generate parasite dynamics post-treatment has been widely validated and used by our group (e.g. (21, 22)) and the approach is being increasingly used by other groups (e.g. (36)). Results are highly robust for both AR-LF and AS-MQ (i.e., partner drugs with different lengths of post-treatment prophylaxis), different levels of transmission intensity, and different levels of drug failures and return an intuitive result (increased failure rate) when very low genetic diversity is simulated. We wish to underline the fact that there are a large number of PK/PD calibrations for AR-LF and AS-MQ in the field; we have chosen the parameterizations here **[SI]** based on our previous work and because their role in the current study is solely to generate plausible profiles of parasite dynamics over time (i.e., figure 1 of (37)) and obtain genetic data with which we can evaluate different methods of molecular correction. We could describe parasite dynamics using other methods (for example, pre-determining a number of clones at a given time and randomly drawing their densities, e.g. (38)) but chose to use a PK/PD model for increased realism, relative simplicity, and to provide the ability for ourselves or other users to calibrate the model to their liking. The crucial part of our methodology is how we calculate detection of microsatellites in blood samples; specifically that it is based on the relative density of alleles in the parasitaemia (and thus dependant on relative clone numbers) and accounts for a "sampling limit" and inherent errors in reading microsatellite lengths. We are confident that while use of different PK/PD parameters would change a given patient's parasite dynamic profile, anything but the most novel parameterization would be unlikely to sufficiently change our results given that a) we simulate 10,000 patients and b) parameters are varied within the model such that a large range of alternative parameterization is already at least partly included in our simulations.

The main practical drawback of the Bayesian algorithm is the need to run a Bayesian analysis. The methodology is published and available (17) but application requires some experience in programming and Bayesian statistics. The analysis is computationally expensive (see **[SI]**) and may be difficult to run on an average personal computer. However, this should not be allowed to be an impediment, given the importance of accurate malaria drug trials, and one solution to this would be for a central body to offer such analyses as a service, or to support application of the algorithm through an internet-based application.

One problem with the microsatellite genotyping approach is its inability to detect low density "minority" clones (the limit here was set to 25%), a problem common to other markers such as the

WHO markers *msp*-1, *msp*-2, *glurp* (the peak height cut-off for ignoring a signal as noise is generally between 10-20% (11)); the slight under-estimates of failure rate produced by Bayesian analyses occurred primarily because the algorithm is unable to correctly identify all low-density recrudescence (Figure 4 and 5) – this reflected minority alleles being missed during amplification and sequencing. There is now considerable interest in using deep-sequenced amplicons as markers, because this method allows detection of alleles at very low frequencies (less than 2% of the frequency of the most frequent allele). We are currently investigating these markers, using a strategy analogous to that described above, to investigate their potential role in molecular correction. Even if they prove accurate and robust, it is likely to be several years before they are validated, a consensus methodology identified and routinely used in trials. Meanwhile it appears that Bayesian analysis of a suite of microsatellite markers does constitute a robust and accurate method for analysis of malaria drug efficacy trials

Author contributions

SJ wrote and conducted the simulations, analysed the results and wrote the first draft of the manuscript

MP wrote the Bayesian algorithm, analysed the results, and edited the manuscript

EMH wrote the simulations and edited the manuscript

KK wrote the simulations and edited the manuscript

IH conceived the project, analysed the results and edited the manuscript

Conflict of interest statement

The authors declare no conflict of interest exists

Funding statement

This research was supported by:

The Medical Research Council (grants G1100522 and MR/L022508/1), the Bill and Melinda Gates Foundation (grant 1032350) and the Malaria Modelling Consortium (grant UWSC9757). MP was supported by the U.S. President's Malaria Initiative.

Role of the funding source(s)

The funding source(s) had no role in study design, collection, analysis or interpretation of data, the writing of the report or the decision to submit the paper for publication.

Meeting(s) where the information has previously been presented

Corresponding author contact information

Sam Jones, Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, United Kingdom, sam.jones@lstmed.ac.uk

Secondary contact

Ian Hastings, Department of Tropical Disease Biology, Liverpool School of Tropical Medicine, Liverpool L3 5QA, United Kingdom, ian.hastings@lstmed.ac.uk

Current affiliations

Acknowledgements

The authors would like to thank Simon Wagstaff and Andrew Bennett of the scientific computing department at the Liverpool School of Tropical Medicine for providing access to the high-performance computing facilities used to generate the results described herein.

We would also like to thank five staff members from the Centers for Disease Control and Prevention for their thoughtful commentary on this manuscript

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. World Health Organization. 2009. Methods for surveillance of antimalarial drug efficacy.
2. World Health Organization. 2016. GLOBAL TECHNICAL STRATEGY FOR MALARIA 2016–2030.
3. World Health Organization MfMV. 2008. Methods and techniques for clinical trials on antimalarial drug efficacy: Genotyping to identify parasite populations.
4. Greenhouse B, Myrick A, Dokomajilar C, Woo JM, Carlson EJ, Rosenthal PJ, Dorsey G. 2006. VALIDATION OF MICROSATELLITE MARKERS FOR USE IN GENOTYPING POLYCLONAL PLASMODIUM FALCIPARUM INFECTIONS. The American journal of tropical medicine and hygiene 75:836-842.
5. Nyachieo A, C VANO, Laurent T, Dujardin JC, D'Alessandro U. 2005. Plasmodium falciparum genotyping by microsatellites as a method to distinguish between recrudescent and new infections. Am J Trop Med Hyg 73:210-3.
6. Greenhouse B, Dokomajilar C, Hubbard A, Rosenthal PJ, Dorsey G. 2007. Impact of Transmission Intensity on the Accuracy of Genotyping To Distinguish Recrudescence from New Infection in Antimalarial Clinical Trials. Antimicrobial Agents and Chemotherapy 51:3096-3103.

7. Anderson TJ, Su XZ, Bockarie M, Lagog M, Day KP. 1999. Twelve microsatellite markers for characterization of Plasmodium falciparum from finger-prick blood samples. Parasitology 119 (Pt 2):113-25.
8. Malvy D, Torrentino-Madamet M, L'Ollivier C, Receveur M-C, Jeddi F, Delhaes L, Piarroux R, Millet P, Pradines B. 2018. Plasmodium falciparum Recrudescence Two Years after Treatment of an Uncomplicated Infection without Return to an Area Where Malaria Is Endemic. Antimicrobial agents and chemotherapy 62:e01892-17.
9. Russo G, L'Episcopia M, Menegon M, Souza SS, Dongho BGD, Vullo V, Lucchi NW, Severini C. 2018. Dihydroartemisinin-piperaquine treatment failure in uncomplicated Plasmodium falciparum malaria case imported from Ethiopia. Infection 46:867-870.
10. Plucinski MM, Huber CS, Akinyi S, Dalton W, Eschete M, Grady K, Silva-Flannery L, Mathison BA, Udhayakumar V, Arguin PM, Barnwell JW. 2014. Novel Mutation in Cytochrome B of Plasmodium falciparum in One of Two Atovaquone-Proguanil Treatment Failures in Travelers Returning From Same Site in Nigeria. Open forum infectious diseases 1:ofu059-ofu059.
11. Messerli C, Hofmann NE, Beck HP, Felger I. 2017. Critical Evaluation of Molecular Monitoring in Malaria Drug Efficacy Trials and Pitfalls of Length-Polymorphic Markers. Antimicrob Agents Chemother 61.
12. Walsh PS, Erlich HA, Higuchi R. 1992. Preferential PCR amplification of alleles: mechanisms and solutions. PCR Methods Appl 1:241-50.
13. Farnert A, Arez AP, Babiker HA, Beck HP, Benito A, Bjorkman A, Bruce MC, Conway DJ, Day KP, Henning L, Mercereau-Puijalon O, Ranford-Cartwright LC, Rubio JM, Snounou G, Walliker D, Zwetyenga J, do Rosario VE. 2001. Genotyping of Plasmodium falciparum infections by PCR: a comparative multicentre study. Trans R Soc Trop Med Hyg 95:225-32.
14. Juliano JJ, Gadalla N, Sutherland CJ, Meshnick SR. 2010. The perils of PCR: can we accurately 'correct' antimalarial trials? Trends Parasitol 26:119-24.
15. Mwangi JM, Omar SA, Ranford-Cartwright LC. 2006. Comparison of microsatellite and antigen-coding loci for differentiating recrudescing Plasmodium falciparum infections from reinfections in Kenya. International Journal for Parasitology 36:329-336.
16. Hwang J, Alemayehu BH, Reithinger R, Tekleyohannes SG, Takele T, Birhanu SG, Demeke L, Hoos D, Melaku Z, Kassa M, Jima D, Malone JL, Nettey H, Green M, Poe A, Akinyi S, Udhayakumar V, Kachur SP, Filler S. 2013. In Vivo Efficacy of Artemether-Lumefantrine and Chloroquine against Plasmodium vivax: A Randomized Open Label Trial in Central Ethiopia. PLOS ONE 8:e63433.
17. Plucinski MM, Morton L, Bushman M, Dimbu PR, Udhayakumar V. 2015. Robust Algorithm for Systematic Classification of Malaria Late Treatment Failures as Recrudescence or Reinfection Using Microsatellite Genotyping. Antimicrobial Agents and Chemotherapy 59:6096-6100.
18. Davlantes E, Dimbu PR, Ferreira CM, Florinda Joao M, Pode D, Felix J, Sanhangala E, Andrade BN, Dos Santos Souza S, Talundzic E, Udhayakumar V, Owens C, Mbounga E, Wiesner L, Halsey ES, Martins JF, Fortes F, Plucinski MM. 2018. Efficacy and safety of artemether-lumefantrine, artesunate-amodiaquine, and dihydroartemisinin-piperaquine for the treatment of uncomplicated Plasmodium falciparum malaria in three provinces in Angola, 2017. Malar J 17:144.
19. Plucinski MM, Dimbu PR, Macaia AP, Ferreira CM, Samutondo C, Quivinja J, Afonso M, Kiniffo R, Mbounga E, Kelley JS, Patel DS, He Y, Talundzic E, Garrett DO, Halsey ES, Udhayakumar V, Ringwald P, Fortes F. 2017. Efficacy of artemether-lumefantrine, artesunate-amodiaquine, and dihydroartemisinin-piperaquine for treatment of uncomplicated Plasmodium falciparum malaria in Angola, 2015. Malar J 16:62.
20. Plucinski MM, Talundzic E, Morton L, Dimbu PR, Macaia AP, Fortes F, Goldman I, Lucchi N, Stennies G, MacArthur JR, Udhayakumar V. 2015. Efficacy of artemether-lumefantrine and

Antimicrobial Agents and Chemotherapy

dihydroartemisinin-piperaquine for treatment of uncomplicated malaria in children in Zaire and Uige Provinces, angola. Antimicrob Agents Chemother 59:437-43.

21. Hodel EM, Kay K, Hayes DJ, Terlouw DJ, Hastings IM. 2014. Optimizing the programmatic deployment of the anti-malarials artemether-lumefantrine and dihydroartemisinin-piperaquine using pharmacological modelling. Malaria Journal 13:1-18.
22. Kay K, Hastings IM. 2013. Improving pharmacokinetic-pharmacodynamic modeling to investigate anti-infective chemotherapy with application to the current generation of antimalarial drugs. PLoS Comput Biol 9:e1003151.
23. Winter K, Hastings IM. 2011. Development, evaluation, and application of an in silico model for antimalarial drug treatment and failure. Antimicrob Agents Chemother 55:3380-92.
24. Anonymous. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
25. H. Wickham. 2016. ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York.
26. Dahal P, Guerin PJ, Price RN, Simpson JA, Stepniewska K. 2019. Evaluating antimalarial efficacy in single-armed and comparative drug trials using competing risk survival analysis: a simulation study. BMC Medical Research Methodology 19:107.
27. World Health Organization. 2016. Malaria Parasite Counting, MALARIA MICROSCOPY STANDARD OPERATING PROCEDURE – MM-SOP-09.
28. Stepniewska K, White NJ. 2008. Pharmacokinetic determinants of the window of selection for antimalarial drug resistance. Antimicrob Agents Chemother 52:1589-96.
29. Garnham PCC. 1966. Malaria Parasites and other Haemosporidia. Blackwell Scientific Publications Ltd., 5, Alfred Street, Oxford.
30. Simpson JA, Watkins ER, Price RN, Aarons L, Kyle DE, White NJ. 2000. Mefloquine Pharmacokinetic-Pharmacodynamic Models: Implications for Dosing and Resistance. Antimicrobial Agents and Chemotherapy 44:3414-3424.
31. Jones S, Kay K, Hodel EM, Chy S, Mbituyumuremyi A, Uwimana A, Menard D, Felger I, Hastings I. 2019. Improving Methods for Analyzing Antimalarial Drug Efficacy Trials: Molecular Correction Based on Length-Polymorphic Markers msp-1, msp-2, and glurp. Antimicrob Agents Chemother 63.
32. Siahaan L. 2018. Laboratory diagnostics of malaria. IOP Conference Series: Earth and Environmental Science 125:012090.
33. T Therneau. 2015. _A Package for Survival Analysis in S_. version 2.38, https://CRAN.R-project.org/package=survival.
34. Alboukadel Kassambara MK. 2018. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.3, https://CRAN.R-project.org/package=survminer.
35. Zweig MH, Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry 39:561-577.
36. Dini S, Zaloumis S, Cao P, Price RN, Fowkes FJI, van der Pluijm RW, McCaw JM, Simpson JA. 2018. Investigating the Efficacy of Triple Artemisinin-Based Combination Therapies for Treating Plasmodium falciparum Malaria Patients Using Mathematical Modeling. Antimicrob Agents Chemother 62.
37. Jaki T, Parry A, Winter K, Hastings I. 2013. Analysing malaria drug trials on a per-individual or per-clone basis: a comparison of methods. Statistics in Medicine 32:3020-3038.
38. Ken-Dror G, Hastings IM. 2016. Markov chain Monte Carlo and expectation maximization approaches for estimation of haplotype frequencies for multiply infected human blood samples. Malar J 15:430.

768
769
770
771
772
773
774
775
776
777
778
779
780
781

782

783 FIGURES

784

785 Figure 1: Failure rate estimates obtained using the match counting algorithm and the Bayesian analysis algorithm for failing AR-LF under low, medium and high
786 transmission scenarios. The true failure rate is denoted in each plot by the horizontal grey line. For the match counting algorithm, the threshold for the number of
787 matching loci at which a recurrence is classified as a recrudescence varies between 2 and 7. For the Bayesian analysis, the cut-off for posterior probability at which a
788 recurrence is classified as a recrudescence varies between ≥0.1 and ≥0.9.

789 Figure 2: Failure rate estimates obtained using the match counting algorithm and the Bayesian analysis algorithm for non-failing AR-LF under low, medium and high
790 transmission scenarios. The true failure rate is denoted in each plot by the horizontal grey line. For the match counting algorithm, the threshold for the number of matches
791 at which a recurrence is classified as a recrudescence varies between 2 and 7. For the Bayesian analysis, the cut-off for posterior probability at which a recurrence is
792 classified as a recrudescence varies between ≥0.1 and ≥0.9.

793 Figure 3: Receiver operator characteristic (ROC) curves showing diagnostic ability of the Bayesian analysis method for 3 scenarios of transmission intensity for non-failing
794 and failing artemether-lumefantrine (AR-LF). ROC curves and area under the roc curve (AUC) are shown for all recrudescence and for high density recrudescence. A high
795 density recrudescence was defined as explained in the main text.

796 Figure 4: Distribution of the posterior probabilities of recrudescence estimated by the Bayesian algorithm for 3 scenarios of transmission intensity for non-failing and failing
797 artemether-lumefantrine (AR-LF). A high density recrudescence was defined as explained in the main text.

798 Figure 5: Contour plot of the posterior probability of recrudescence estimated by Bayesian algorithm as a function of the density of recrudescent clones (i.e., the
799 proportion of the recrudescent clones in the total recurrent infection biomass) in the initial sample and the recurrent sample. This plot is the combined data of all 6
800 scenarios modelled for artemether-lumefantrine (AR-LF). Each contour line indicates the posterior probability of recrudescence and the area between the lines the number
801 of recurrent infections in the population with those posterior probabilities.

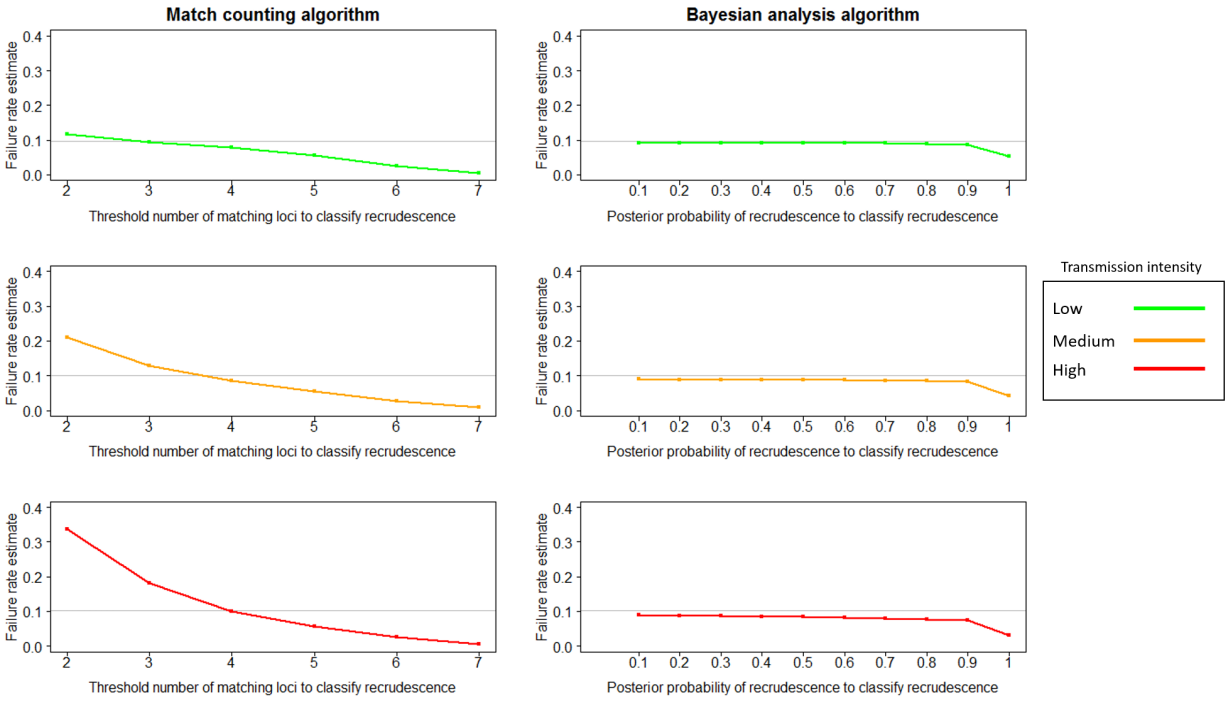
802

803

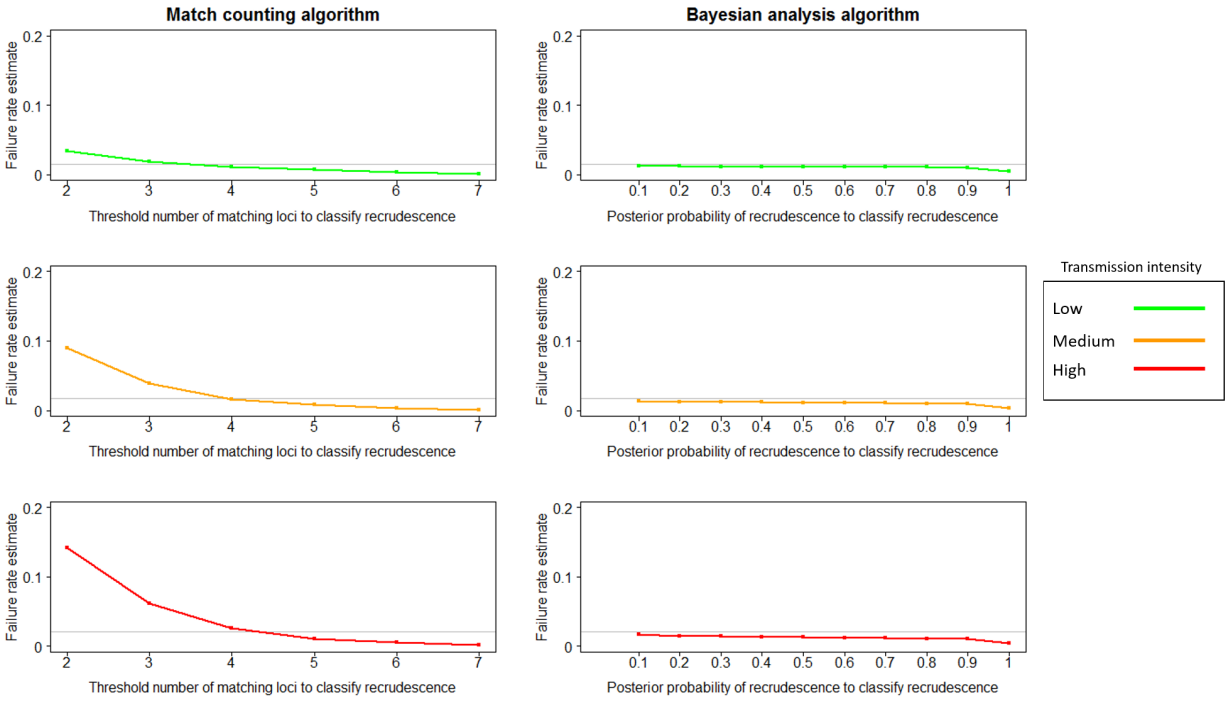
804

805

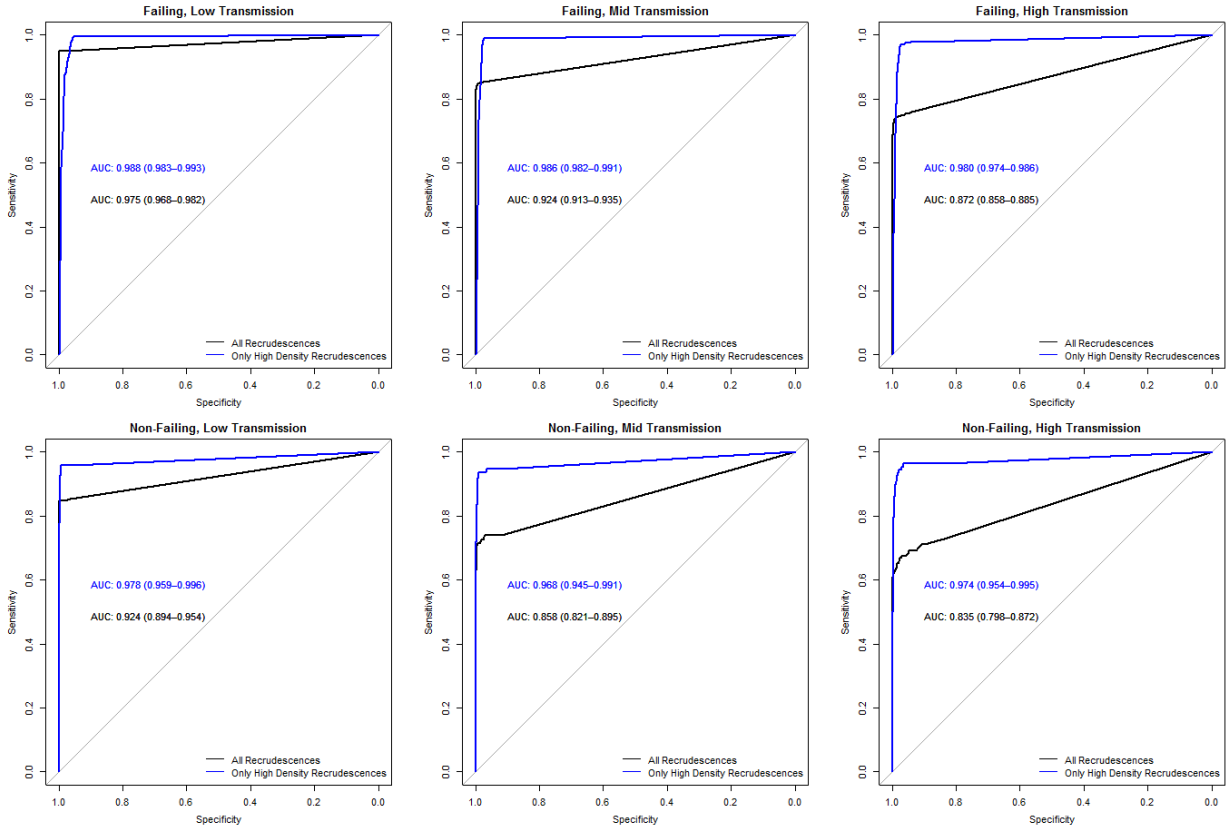
806



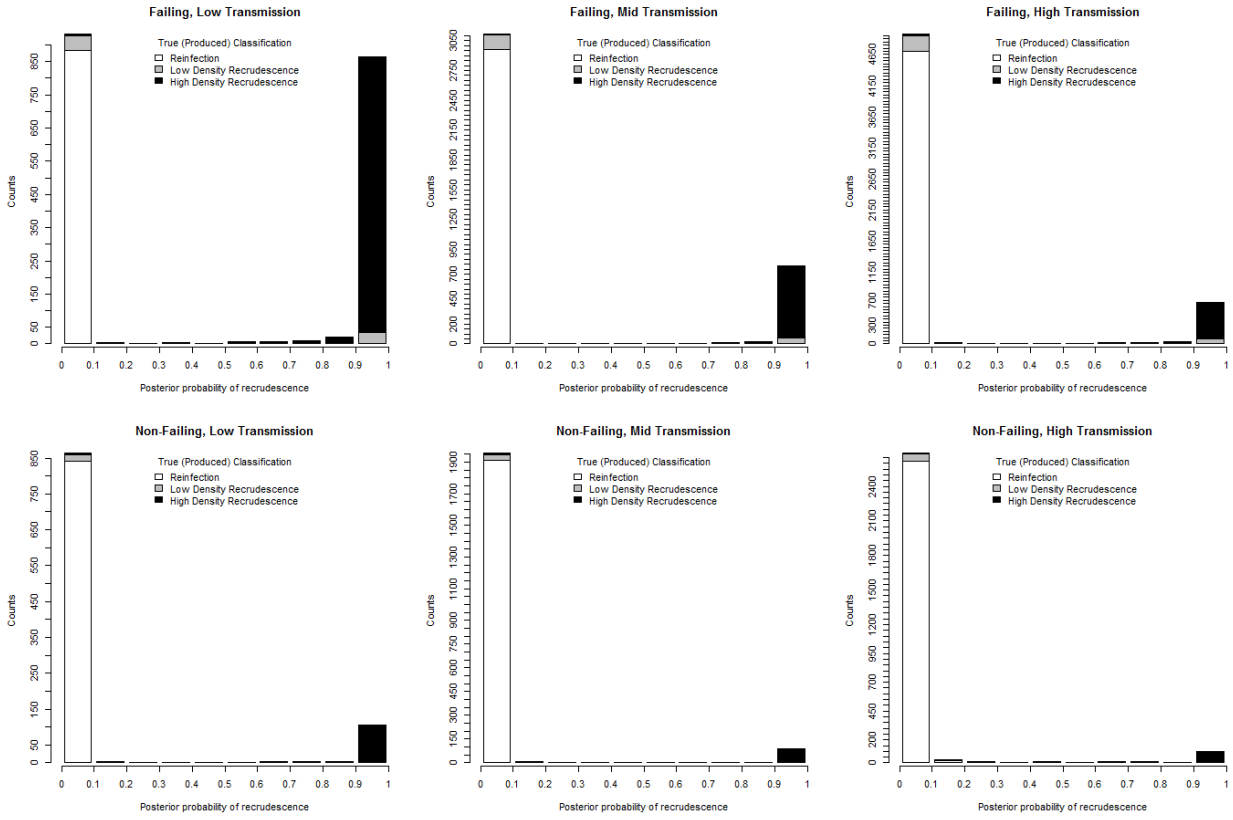
807



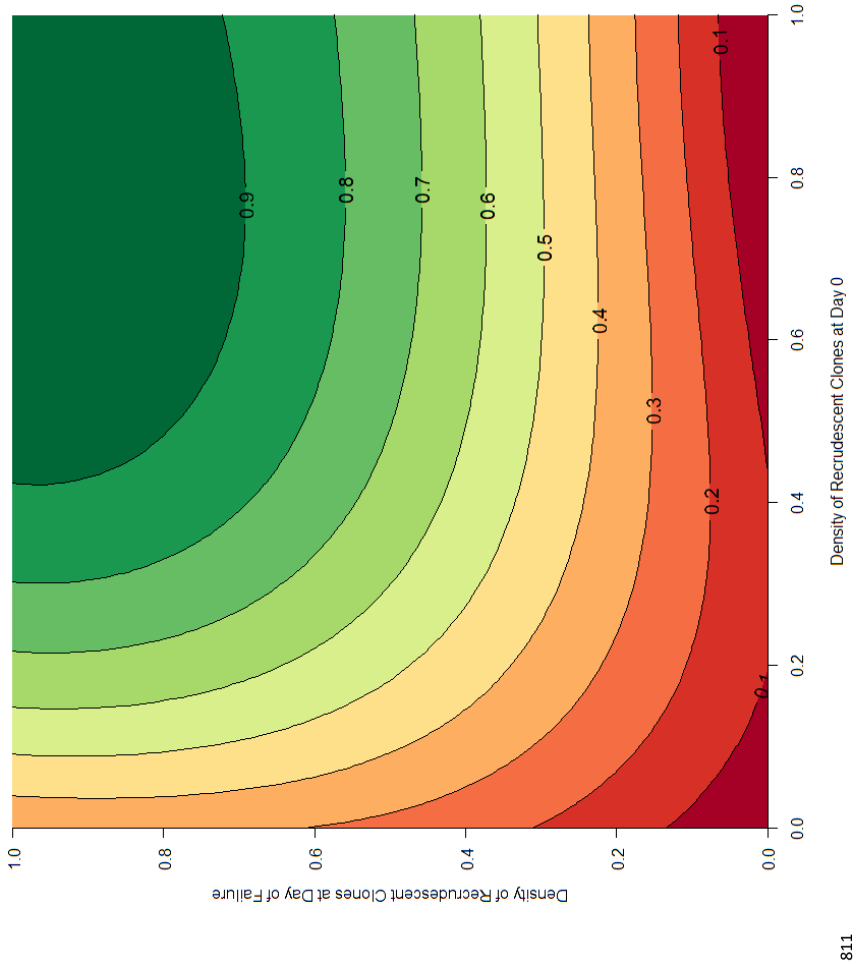
808



809



810



811